A framework for assessing empirical approaches to moral philosophy

Daniel Thomson

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

University of Canterbury

Abstract

In this thesis I examine the question 'What are the implications of our growing scientific understanding of moral phenomenon for ethics?' I assess what moral philosophy can gain, if anything, from the rapidly growing body of literature on the evolutionary genealogy, psychology, and biology of morality. Historically, there have tended to be two kinds of responses to such questions: either great enthusiasm about the potential to revolutionise ethics followed by dramatic conclusions without adequate rationale; or empirical considerations are rejected as irrelevant by invoking 'Hume's law' (that one cannot derive normative conclusions from descriptive facts) and not examined further.

I argue for an approach that takes a middle ground; that our growing scientific understanding may have implications for a number of debates in moral philosophy, but that at the same time, there are few conclusions that are obvious or straightforward. Due diligence must be given to philosophical analysis to adequately assess relevant empirical research. There is in general no principle that allows us to determine whether empirical research is relevant to its related areas of philosophy, thus I argue that any such evaluation, at least initially, must be done on a case by case basis. I examine a number of case studies of different authors who try to derive implications for ethics from empirical research into morality. From looking at these cases I also evaluate what can be learned from these attempts and present the findings as a framework that can be used when assessing the philosophical merits of empirical approaches to moral philosophy.

Contents

Acknowledgements

Over the course of writing this thesis I have had several supervisors who all deserve thanks. My first supervisor, Graham Macdonald, also gave the first philosophy lecture I ever attended and has continued to provide inspiration since that day. Derek Browne, Michael-John Turp, and Doug Campbell all provided supervision at various points over the years and gave valuable assistance, discussion, and encouragement. Carolyn Mason, who took over as my senior supervisor, deserves special thanks for her guidance, kindness, and endless patience throughout the thesis. Carolyn's feedback was always helpful and she somehow managed to provide consistent encouragement (prodding!) whilst still being unwaveringly supportive.

I also owe thanks to the friends, office mates, work colleagues, and fellow students who are too numerous to name individually but who have all helped, supported, and made things better in any number of ways over the years while writing the thesis.

Finally, I would like to thank all of my family for their endless support and especially my parents, Robyn and Andrew, who deserve more thanks than I can give them.

Chapter 1 Introduction

In this thesis I examine what philosophical ethics or moral philosophy can gain, if anything, from the growing bodies of literature on the evolutionary genealogy, psychology, and biology of morality. In recent years a number of attempts have been made that argue that such research has relevance for ethics or that one can derive ethical conclusions from it. This empirical research includes work in disciplines such as evolutionary biology, evolutionary psychology and its predecessor sociobiology, various branches of psychology, neuroscience, anthropology, and others. I look at whether such attempts have been successful and examine what makes successful attempts so, and where unsuccessful attempts fall down. I analyse what we can learn from these attempts and provide a framework for assessing or developing an argument from empirical ethics.

Historically, there have tended to be two kinds of responses to attempts to draw ethical conclusions from empirical discoveries: either great enthusiasm followed by jumping to conclusions without adequate philosophical analysis; or such attempts are rejected as irrelevant by invoking considerations such as Hume's law (that one cannot derive normative conclusions from descriptive facts) and not examined further.

This work is for the most part, descriptive in nature. It is often claimed therefore that it has little relevance to moral philosophy, and any suggestion of significance is usually warded off by repeating the catchphrase derived from Hume that 'one cannot derive an ought from an is'. This catch phrase has its origins in the following passage from Hume's *A treatise of human nature*:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd;

> and at the same time that a reason shou'd be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.[1]

A commonly accepted reading of this passage is that no evaluative or normative conclusion can be inferred from any set of purely descriptive or factual premises. This general view was termed 'Hume's Law' by R. M Hare[2] and has been used as one of the most common responses to claims that scientific research into morality can have any significance for moral philosophy. I take this consideration to be a poor reason to abstain from examining empirical work into morality for its philosophical merits or insights for two reasons.

Firstly, whether this interpretation is correct is contentious, and much philosophical effort has gone in to both attempting to decipher precisely what point Hume was trying to convey and further, whether some form of Hume's Law is actually true and if it is, in what manner it is true. In an attempt to move forward and not belabour philosophical discussions that have already been overworked, I take the following conservative position on the is/ought debate[3]: what Hume is establishing is a point about deductive arguments; that if a moral 'ought' occurs in the conclusion of some deductive argument, but not anywhere in the premises of the argument, the argument is *invalid.*[4] This point about logical deduction does not imply that 'normative conclusions cannot be inferred from factual or descriptive premises', only that deductive arguments containing oughts in their conclusions must have an ought in one of the premises for the deduction be valid.

Secondly and more importantly, it is not true that moral philosophy is interested in only those things that fall cleanly on the normative or 'ought' side of the divide between 'is' and 'ought'. There are many

[1] David Hume, *A treatise of human nature*, Book 3, sec. 1, part 1, p. 334.
[2] R. M. Hare, 'Universalisability', p. 303.
[3] It is worth noting that the is/ought "gap" is sometimes mislabeled as the 'naturalistic fallacy', but this is a misnomer as the term 'naturalistic fallacy' was coined by G. E. Moore in his *Principia ethica* for a *prima facie* similar, but nevertheless different, point about the *semantics* of moral terms; the 'naturalistic fallacy' refers to the idea that it is mistake to attempt to define moral terms such as 'good' in terms of natural properties such as 'desirable'.
[4] This is roughly the position taken by Charles Pigden. See 'Snare's puzzle/Hume's purpose: Non-cognitivism and what Hume was really up to with no-ought-from-is'.

subtle distinctions involved in evaluating which moral judgments should be made, whether such judgments are true, or correct, or justified, or 'appropriate' in some other sense. These subsequent questions are questions about the *nature* of moral debate – questions that are usually considered to fall under the umbrella of metaethics. These questions nevertheless form an integral part of the endeavour of moral philosophy.

Metaethics is sometimes defined as simply 'the study of the nature of morality', but this can be misleading, as this definition makes it sound as though it is the kind of thing science can directly investigate – that metaethics might be a solely descriptive domain. But the questions metaethics traditionally deals with are not obviously amenable to being dealt with in an empirical manner. When philosophers inquire about the nature of morality, they are asking questions such as the following:

- When we make moral judgments, are we stating facts? If so, what kind of facts are they? Or, are we instead expressing our attitudes or desires?
- Can moral judgments be true or false? If not, can they be more or less 'correct' or 'appropriate'? How are they justified? *Are* they ever justified?
- How do we come to *know* that they are true or correct or justified? Can we have moral knowledge? What kind of epistemology is appropriate in the moral domain?
- Is there such a thing as objectivity in ethics? Or is ethics fundamentally subjective or relative to some agent or group of agents? Can we say what the objective/subjective distinction amounts to in ethics?
- Is there such a thing as moral progress? What does it consist in, and how do we know when it has been made?

So it is clear that the idea that there is an unassailable gap between normative and the descriptive disciplines and therefore empirical research into morality is not relevant to moral philosophy is too simplistic to be used as a reply to reject such considerations outright.[5] There is no principle that can be applied in general to the question of whether any given item of empirical research into morality

[5] For a further discussion of why we should reject philosophy's historic default position of claiming irrelevance for such considerations, Oliver Curry's 'Who's afraid of the naturalistic fallacy?' is a comprehensive discussion. Curry argues that few philosophers have been willing to pursue the 'naturalistic approach' due to a range of things that have been called the 'naturalistic fallacy' including Hume's Law and Moore's naturalistic fallacy. Curry discusses several related concerns that have been given this name and demonstrates none of them pose significant problems for attempting to integrate empirical work with philosophical ethics.

will be of *normative* moral significance or of significance for metaethics or any other area of philosophical ethics. Nevertheless, we must proceed cautiously, as drawing conclusions from biology and the empirical sciences for ethics is not easy; as Philip Kitcher has described it "the relation between biology and ethics has been an alluring swamp in which any number of scholars have floundered."[6]

Often in interdisciplinary contexts, usage of terminology differs between disciplines. Useful engagement between disciplines requires understanding what each discipline is talking about, so before moving on to chapter 1, I will provide an account of the ways in which 'morality' is being used so that it is clear what the phenomenon to be explained is and to disambiguate where possible how certain researchers from differing disciplines are using the term. The characterizations that I give will be general, and non-committal on some points: indeed, the term 'morality' without context probably does not have a determinate enough meaning to be able to give a non-controversial list of necessary and sufficient conditions that all would agree with. Nevertheless, I hope to provide a general account that most (including philosophers) could agree is a representative description. This description of morality will hopefully be neutral on most normative and meta-ethical questions, as the intention here is to give a clear picture of what the phenomenon appears to be rather than an analysis that shows what is 'really going on' or settles any justificatory questions. As there are many different uses of the term 'morality' it is useful to start by having a clear picture of what these uses are and which meanings are intended in certain contexts.

'Morality' is a word not frequently used outside of moral philosophy, and when it is, it is often used differently to its usage within philosophical circles. Dictionary definitions typically list two senses in which the word is used: a normative and descriptive sense. When used descriptively, 'morality' usually refers to a code of conduct or principles concerning the assessment of behavior put forward by a group or society or to assess one's own behavior. This descriptive usage is what is generally intended when

[6] Philip Kitcher, 'Biology and ethics', p. 163.

biologists, anthropologists, psychologists, sociobiologists and so on, talk about the 'evolution of morality'. However, strictly speaking morality is the output of a moral psychology, and it is this particular psychology that is what evolved rather than 'morality' itself.

The normative usage of 'morality' is that which philosophers often concern themselves with. When a moral philosopher asks 'is act *X* moral?' (or questions about this question) they are asking is *X* in accordance with some idealized code of conduct that would be put forward by all rational beings (or if not rational, whatever method of assessment they argue is the correct one). Thus moral philosophy concerns itself with not only the question 'is *X* moral?' but also 'is the code of conduct which we are asking if *X* is in accordance with the *correct* one?' (or sound or true or some other kind of similar assessment or agreement with it).

'Moral judgments' are the assessments or evaluations that individuals make about whether something is in accordance with the code of conduct or shared rules of how to live. They appear to be a kind of mental event – an assessment of a past, present, or future situation, usually resulting in some kind of linguistic utterance or conclusion thought to oneself. They take as their subject matter the conduct of people in interpersonal relations. A moral judgment can be about one's own interactions with others, or between about the conduct and interaction of third-parties. They are most likely to be about (but certainly not limited to): negative appraisals of acts of harming others, assessments of reciprocity and fairness, requirements concerning behaviour in a matter befitting one's status in a social hierarchy, and rules about 'bodily matters' – things such as food, sex, hygiene, appropriate practices concerning bodily waste and so on.[7] More generally, morality is often (especially in philosophy) taken to be about the sum of all these considerations: that is, how one ought to live or what one should do all things rationally considered. Moral judgments also often involve a component of desert. That is they involve coming to positive or negative conclusions that people *deserve* a particular treatment: a punishment or reward as a result of their conduct. The notion of being

[7] List adapted from Richard Joyce, *The evolution of morality*, p. 65.

deserving of a reward or punishment can be both external (that is judging that others are *deserving* of something) or internal to the individual making a moral judgment, in the form of a moral conscience.

The precise kind of utterance that verbalizations of moral judgments are, is a contentious issue in moral philosophy. One point upon which views typically diverge is whether moral judgments assert beliefs about facts (things that can be true or false) or express attitudes (such as disapproval, disgust, acceptance of some standard and so on). These two viewpoints fall roughly under the terms 'cognitivism' (that moral statements have cognitive content: they are beliefs) and 'non-cognitivism' (the denial of cognitivism). In describing this aspect of moral judgments, I mean only to provide an account of the features that moral judgments *appear* to have; active philosophical research on the issue is ongoing with well entrenched positions. Firstly, there looks to be much that is correct about the cognitive picture of moral judgments. Moral statements usually take the form of a subject-predicate sentence such as "Stealing is wrong" or "Hitler is evil". Such sentences contain a predicate applying what seems to be a property (wrongness) to the subject (stealing). In non-moral discourse it is generally accepted that sentences of this form, such as "the fridge is white" assert beliefs. So, *prima facie* the grammatical form of moral judgments points towards their being assertions of beliefs. A second point in favour of the cognitivist is that we do not simply put forward moral judgments, we also *dispute* them. Arguing that something is right or wrong is a common occurrence, and such discourse makes the most sense if what people are arguing about are facts, things that can actually be true or false. Thus, a good first approximation is that moral judgments appear to express beliefs.

In many contexts in which moral judgments are put forward however, they can also appear to have a non-cognitive element: they express some kind of attitude. This feature of judgments can be brought out clearly with the following simple argument.[8] Consider the following sentence:

(1) "Racial discrimination is wrong"

[8] This form of argument is developed in greater depth in Richard Joyce, see *ibid.*, pp. 53-57.

Like the examples in the previous paragraph, if one heard this phrase uttered, it would perhaps most naturally be taken to be expressing a belief. However, if someone uttered the following:

(2) "Racial discrimination is wrong, but I do not disapprove of it"

It is likely that we will be perplexed as to what they mean. Such a sentence appears to contain some kind of contradiction, even if it is not apparent where precisely it lies. Upon hearing this sentence, our initial reaction might be to try to ascertain what was abnormal about the context. Was the speaker joking? Was it said in a sarcastic tone? Do they think that everyone else thinks that racial discrimination is wrong, but they themselves dissent from this view? Are they simply confused, or trying to confuse others? The point of this is that when we try to interpret what (2) means, we look for reasons why it might not be a genuinely made moral judgment. Thus, if in (2), "but I do not disapprove of it" is enough to undermine the seriousness of what precedes it ("Racial discrimination is wrong"), then the most straightforward explanation is that the "Racial discrimination is wrong" part of (2), functions (at least in part) to express disapproval in addition to appearing to assert a fact.

While this result is a departure from the usual exclusive endorsement of only one of cognitivism or non-cognitivism, there does not seem to be anything problematic with the thought that some linguistic utterances can express more than one mental state at a time. For example calling someone a 'bible basher' indicates that you have a belief that they are a Christian who tends to evangelize, but this is not *all* that you express by using the term: it is usually also an expression of disdain for the aggressive tactics of proselytizing used and an aversion to the message in general (the person who was called a bible basher is most unlikely to describe themselves in this way!).

A final feature of moral judgments is that they appear to have a certain kind of force, a special importance, to them. By this it is meant that in some sense moral judgments are made with the intention of being *inescapable* and that they purport to have a kind of *authority* that provides genuine deliberative weight. The inescapability of moral judgments amounts to the fact that they are intended to apply whether the subject of them agrees or disagrees with the judgment. It is important to note

that whether these considerations are taken any notice of is another issue: the claim is not that moral judgments are necessarily motivating, only that they are made with the intention of not being able to be ignored as considerations. In this respect, moral judgments often seem to be like categorical imperatives – an imperative that's legitimacy does not depend on some goal aimed at by its target. A hypothetical imperative is the opposite – one which does depend on the agent in question's goals. If for example I say "Have some cake", there is generally a tacit conditional of "unless you're not hungry" or something similar implied: the imperative here is 'escapable' in some sense. By offering cake, I generally do not want to advance an imperative that should be followed regardless of your wants or desires. Thus, "Have some cake" is hypothetical in a way that "Do not murder people" is not: there are no conditions such as "unless you enjoy murdering people" that one could have that would invalidate the imperative. That moral judgments have the form of categorical imperatives is a relatively uncontroversial claim as long as it is remembered that this is only a description of a feature that moral judgments appear to have. That they just do appear to be stated as imperatives that are inescapable should not be confused with the claim that the moral judgments actually are inescapable or that their categorical form provides a fundamental basis or justification for morality.

In addition to inescapability, moral judgments gain their practical weight from the *authority* that they purport to have. Moral authority is another controversial topic in ethics, but it does seem to be true that moral judgments are intended to carry a certain kind of weight or importance that is greater than other kinds of judgments such as judgments of transgressions of conventions or prudential requirements. What the source of such authority is does not appear to be obvious or epistemically accessible to moral judgers, but it is a robust and consistent feature that moral judgments appear to have; studies show that children as young as 3 attribute a special kind of authority to moral judgments.[9]

[9] In one study, children are asked whether breaking a convention – a boy wearing a dress to school – is "ok", and most respond negatively, but if asked "would it be alright if the teacher were to say it's ok?" then most respond that it would be. For a moral judgment such as whether punching another student is ok, respondents said that

To summarize, moral judgments to appear to be:

- Mental evaluations of situations concerning interpersonal relations. These evaluations are often subsequently expressed in linguistic utterances.
- Their subject matter is mainly issues of harm, fairness, justice, social status, and appropriate practices concerning 'bodily matters'.
- They often imply a concept of desert: deserving rewards or punishments.
- Statements of judgments (the linguistic utterances) have features that make them *appear* to express both beliefs and attitudes.
- Moral judgments purport to have a special kind of inescapability and authority.

1.1 Thesis Structure

The structure of this thesis is summed up as follows:

1. Disciplines outside of moral philosophy have been interested in the phenomena of morality and have been producing research concerning morality's nature, function, operation, and place in the world.
2. The interactions of moral philosophers and these researchers from outside philosophy generally has been of two kinds in the past: either much of this research has been routinely ignored or deemed irrelevant because of Hume's Law or other related considerations, or philosophical arguments and conclusions have been advanced (often by those doing the empirical research themselves) as being revolutionary for moral philosophy, but have arrived at their conclusions with scarce or inadequate philosophical analysis done and consequently have been ignored by mainstream moral philosophy.

it was never ok, even if the teacher were to say it was. See Judith G. Smetana and Judith L. Braeges, 'The development of toddlers' moral and conventional judgments.'

3. There is no currently existing principle for deciding in general how in a particular case to deal with such things.
4. So, the best we can do is look at a range of different attempts to say something important or philosophically interesting based on empirical research into morality, and see:
    a. What if anything there is to these claims about the relevance of empirical research into morality, and in the cases where the arguments are successful, what the philosophical implications are, and
    b. In examining these case studies, what we can learn about how to assess the impact of empirical research into morality: what lessons, organizing principles, strategies, pitfalls, rules, and tests that can be applied to instances of philosophical arguments utilizing empirical approaches to morality.
5. The result of this is a framework for assessing empirical approaches to moral philosophy and a better understanding of whether such approaches to moral philosophy are likely to be successful or fruitful.

The intended audience of the framework developed in point 5 above, are nonphilosophers working on interdisciplinary research involving philosophy or having philosophical conclusions. In their paper 'Interdisciplinary collaboration in philosophy', Andrew Higgins and Alexis Dyschkant have argued that philosophers should take a more collaborative approach towards other academic disciplines. This integrationist position advocates supplementing philosophy's methodology with the methods and tools of the sciences and other disciplines.[10] Higgins and Dyschkant argue that philosophers actively should communicate and collaborate with nonphilosophers on research and that such cooperation would lead to significant benefits for philosophers and progress of philosophy.

[10] Andrew Higgens, Alexis Dyschkant, 'Interdisciplinary collaboration in philosophy'. See also Andrea Polonioli, 'New Issues for New Methods: Ethical and Editorial Challenges for an Experimental Philosophy'.

However, not all collaboration of this kind has been successful due to philosophers' lack of experience and understanding of the methodology of other disciplines. They note that "too many experimentalists have failed to produce philosophically significant work, either because they have insufficiently integrated themselves into the research communities they aim to emulate or because they have not fully taken advantage of the distinctive methods of philosophy."[11]

To address this deficit, they argue that philosophers should work with nonphilosophers in those complementary disciplines. They also provide some guidelines for philosophers working in this interdisciplinary space in an attempt to avoid common missteps and violations of the methodological norms of those other disciplines. The framework in this thesis is produced in a similar spirit and its intended audience is the other side of such interdisciplinary collaborations. It aims to provide guidance for nonphilosophers collaborating with philosophers to make forays into philosophical debates or for those not currently collaborating with philosophers who have become excited by the potential implications for philosophy of their empirical findings.

In the first part of the thesis, consisting of chapters 2, 3, and 4, I examine one area of empirical and theoretical research which it has been argued has significant implications for philosophical ethics: the evolutionary origins of morality. In chapter 2 I introduce the recent research describing the evolutionary genealogy of morality to show that such an evolutionary story has much plausibility.

In chapter 3 I turn to discuss the potential implications this growing understanding of the evolution of morality has for philosophical ethics. I look briefly at the attempts by the biologist E. O. Wilson to draw ethical conclusions from the evolutionary origins of human nature and sociality. When first published, Wilson's work on human morality and his claims about ethics were highly controversial. He argues that evolutionary biology can shed light on a number of issues: on the metaphysics of morality, on what he calls the 'problem of altruism', on biological constraints on what we ought to do, and on the 'naturalness' of various human behaviours and consequently their ethical status among other issues.

[11] Andrew Higgens, Alexis Dyschkant, 'Interdisciplinary collaboration in philosophy', p. 373.

Wilson's arguments are suggestive of a number of different ways in which an understanding of our biological origins may provide novel insights into ethics. While suggestive, I conclude Wilson's arguments are either unsuccessful or underdeveloped, and can be used to highlight a number of common difficulties or complications in drawing ethical conclusions from evolutionary facts. His work shows that useful philosophical conclusions do not simply fall out of the facts about morality's evolutionary genealogy and require argument and engagement with the philosophical literature to make progress.

Secondly, and in more detail, I examine the work of Richard Joyce and Sharon Street who provide more sophisticated and in-depth attempts at drawing implications for philosophical ethics from the evolution of morality. Joyce argues that because we have evolved to have certain moral beliefs due to the evolutionary advantage they gave, moral truth may not have played any part in the genealogy of morality and consequently our moral beliefs are in danger of having their justification undermined. In contrast to Wilson, Joyce locates his argument in the wider meta-ethical context and addresses the philosophical difficulties involved in arguing for his position, and his evolutionary argument raises novel considerations for meta-ethical debates about moral scepticism.

In a related area to Joyce's argument, Sharon Street puts forward a Dilemma for the Moral Realist that aims to challenge them to choose between dropping their commitment to a scientifically defeasible account of the evolution of morality and rejecting their theoretical commitment to moral realism. According to Street's argument, if the Moral Realist wishes to avoid the latter choice, they must argue that somehow the independent moral truth that Realists are committed to was involved in the evolutionary genealogy of morality. Arguably, no realists have adequately met this challenge, and indeed it is hard to see how they would go about doing so. If the Moral Realist wishes to reject that independent moral truth played a role in the evolution of morality, then they must either accept anti-realism or claim that a coincidence of fantastic proportions has occurred and our evolved moral psychology just happens by chance to be identical to the independent moral reality.

In Chapter 4 I examine the outcome of these attempts to put forward philosophical arguments based on the evolution of morality and assess what we can learn from these attempts. I argue that ultimately Wilson's attempts are unsuccessful, but that his work is instructive in a number of ways and highlights a number of pitfalls in such attempts. Richard Joyce's arguments are more successful and provide a good model of how to integrate empirical and ethical research. While Joyce's conclusions are somewhat controversial, and he establishes less than he initially sets out to, he nevertheless establishes important and noteworthy results. Sharon Street's argument is more successful in establishing what she sets out to achieve, namely showing that evaluative realism that has a commitment to mind-independence is untenable. This is a somewhat limited conclusion, but is nonetheless important and provides a new and sound argument against some forms of moral realism. It pushes the meta-ethical debates in the direction of a more moderate picture of the metaphysics of morality and this is an important contribution. When taken together, Street and Joyce's arguments provide a picture of morality in which the metaphysics of morality is limited to certain options.

I examine where each of Wilson, Joyce, and Street were successful, where they went wrong, and what we can learn from their attempts. From looking at the manner of successes and failures of these attempts and the lessons we can learn from their missteps, I begin constructing a framework which can be useful in assessing arguments of this empirical or evolutionary sort.

In Chapter 5 I turn to examine the implications of research in psychology for ethics. I give an overview of recent work done on moral psychology that is thought to be of relevance to moral philosophy, including models of how moral judgment are made by Joshua Greene, Marc Hauser, Jonathan Haidt, and Shaun Nichols. I present a combined tentative model of this work, and discuss the philosophical work of Nichols who argues for certain theses about moral rationalism and emotivism.

In Chapter 6 I discuss how empirical considerations about how moral judgments are made may or may not have implications for philosophical ethics. There are however limits to the usefulness of discussing in the abstract how important the implications of moral psychology for ethics are, and thus following

these general remarks I examine two different attempts to draw ethical conclusions from research done in moral psychology and in chapter 7, social psychology.

The first that I look at is from the work of Shaun Nichols, who claims that psychological findings undermine 'moral rationalism'—the idea that moral judgments are closely connected in some way to rationality. Nichols thinks that findings concerning psychopaths, people who appear to be unmotivated by moral considerations but have intact rational capacities, show that various rationalist theses are false. He argues that psychopaths provide a counter-example to empirical rationalism; the idea that moral judgments are produced by rational faculties. I evaluate this argument and conclude that while the form of rationalism he argues for is in principle a hypothesis that could be tractable to empirical methods, there are a number of problems with using the evidence Nichols cites. Nichols also argues that conceptual rationalism, the idea that it is part of our concept of a morality that moral requirements are rational requirements, is shown to be mistaken by data about people's concept of a psychopath. I argue that this attempt is also unsuccessful due to a number of difficulties involved in assessing the content of concepts.

Secondly, I examine an argument put forward by John Doris and Stephen Stich, who claim that findings in social psychology undermine the assumptions of virtue ethics. They argue that various 'situationist' findings show that people do not have robust dispositions to act virtuously in the sense that virtue ethics assumes people have. Thus, their conclusion is that psychological data shows virtue ethics to be an untenable theory. I argue that their arguments are unsuccessful for two reasons. Firstly, their interpretation of the findings of 'situationist' psychology is too extreme, and they generalise too far from these results. Secondly there are doubts about whether the conception of virtue they target is the same conception as that used by virtue ethicists.

In chapter 7 I repeat the analysis done in chapter 4 but for the arguments presented in chapters 5 and 6. I examine what we can learn from attempts to argue for philosophical conclusions based on empirical research into morality by Shaun Nichols, Adina Roskies, John Doris and Stephen Stich. I look

at the successes and failures of the presented arguments, and analyse what we can learn from these attempts about what a good and bad argument for ethics from empirical research looks like. I draw from this analysis the second half of the framework.

In chapter 8 I present my conclusions about the success of the case studies in drawing conclusions for moral philosophy from empirical research and I present the framework to assist in assessing and developing empirical approaches to morality.

Chapter 2 Evolutionary origins of morality

2.1 The evolution of morality

The idea that evolution has may have pernicious implications for morality dates back to the time when the theory of evolution was itself developed. There have been many attempts at using evolution to undermine the authority of morality or to reveal that "morality is a collective illusion foisted upon us by our genes."[12] Often the mainstream philosophical response to these arguments is to dismiss them as naïve or misled. Nevertheless, despite the recognition that many of these arguments have been poor, there are philosophers and researchers that are intimately familiar with the details of the evolutionary origins of morality who think that an understanding of its origins *must* still somehow inform us of useful information about morality's nature and metaphysics. Indeed, the more familiar one becomes with the evolutionary genealogy of morality the harder it becomes to see how it cannot have some impact on the kind of thing we think morality is. If moral philosophy is concerned with understanding morality, then asking key analysis questions of 'why' at different levels of explanation; including the evolutionary or 'ultimate'[13] levels, cannot help but be informative. Philosophy often progresses when scientific advances make some philosophical questions redundant or through so called 'disciplinary speciation' where new disciplines emerge from work in philosophy or another discipline such as physics, psychology, linguistics, and economics and so on. As David Chalmers notes, "these fields have sprung forth as tools have been developed to address questions more precisely and more decisively…when we develop methods for conclusively answering philosophical questions, those methods come to constitute a new field and the questions are no longer deemed philosophical."[14] Therefore in this section I briefly attempt to make as plausible as possible the case for evolution playing a central role in the origins of morality.

[12] Michael Ruse, *Taking Darwin seriously*, 1986, p. 253.
[13] To borrow Nikolaas Tinbergen's terminology for evolutionary explanations in biology, originally from 'On aims and methods of ethology'.
[14] David Chalmers, 'Why isn't there more progress in philosophy?', pp. 20-21.

2.2 The evolution of cooperation

I begin by presenting a brief overview of the processes that gave rise to cooperation and sociality in organisms, capacities that are widely held to be the pre-cursors to our moral faculties. I then discuss how various processes may have resulted in the transition from cooperation and group living to more full-fledged systems of norm-based living and morality. While much of the theory of the emergence of cooperation and sociality in organisms is well accepted, the more cognitive and culturally infused elements of the evolution of morality are more conjectural and tentative. Nevertheless, there is a growing sense that *something* along such lines must be correct: evolutionary processes must have played a large role in shaping the psychological capacities humans use to judge morally and live as social creatures.[15]

Any evolutionary explanation is one that explains the existence of a trait by a process of variation and selection. This process consists of small heritable variations in structure having an effect on the phenotype of an organism, which in turn influences its reproductive success or 'fitness'. Those variations in genotype which result in phenotypes that improve an organism's fitness become more common in successive generations. Adaptations are the cumulative outcome of many iterations of this process: they are the features of organisms that have an identifiable structure, function, use, or form[16] (this includes both adaptations of the 'engineering' type and so called 'selection-byproduct' adaptations[17]). Evolutionary pressures that persist through many generations make certain variations

[15] As William Fitzpatrick notes in 'Morality and evolutionary biology' this has been a growing theme in both popular and academic writing, echoing themes in E. O. Wilson's *Sociobiology* (although certainly Wilson is not the first to make such claims).

[16] This is somewhat ambiguous: there is continuing debate as to whether the term 'adaptation' should be reserved for features that have an identifiable adaptive function or purpose that has been selected for (sometimes termed *engineering* type adaptations), or whether it should be used more broadly to refer to any feature or trait that is the result of evolution (called a *selection-product* type adaptation). For further discussion see Elizabeth Lloyd, 'Units and levels of selection: an anatomy of the units of selection debates'.

[17] Note this distinction and similar ways in which adaptation and selection can diverge or be present separate to each other is also discussed below in section 2.3.3 'From sociality to morality – epistemic difficulties'.

in the same 'direction' consistently fitness enhancing. The many iterations of this process of reproduction and selection 'accumulate' and eventually result in identifiable traits.

The thought of adaptations generally brings to mind images of structural or physiological features of organisms such as beaks of a certain shape or size, organs such as gills, specific colours or markings and so on. There is however no reason why adaptations are limited to such physiological features: many adaptations are behaviours or mental capacities or structures that mediate adaptive behaviour. Morality is the product of traits of this last kind; it is a social phenomenon that is the output of a number of interacting and interrelated human faculties, each of which in part, has evolutionary origins. To begin answering questions therefore about morality's phylogeny, we need to examine its particular parts, and look at how these elements might have first emerged and only then how they may have developed into more complex phenomena. Accordingly, I shall begin by looking at the simplest forms of helping behaviour that emerged.

Helping behaviour simply means behaviour that one organism produces that has beneficial effects for the fitness of another organism. There are no other requirements: the behaviour does not have to be consciously performed, it does not have to be performed with the intent to be helpful, and there is no requirement that it be 'altruistic' in the everyday sense of the word.[18] By looking at the most basic systems of regulation of social behaviour, we can start to theorize how more complex systems of cooperative social behaviour emerge. That such basic processes and systems that are part of our phylogeny should still have an influence on our present moral faculties should not be surprising. Evolution works by modification of what is already present; new structures or systems come about through modification of previous ones. In the case of morality, we should expect the same to happen: more complex forms of mental faculties that allow for social living are modifications and extensions of prior, simpler forms of mental faculties for social living. The continuity of evolutionary processes

[18] The term 'altruistic' is often used with one meaning in biology and another in other disciplines, which can result in confusion when the contexts intersect. In biology 'altruistic behaviour' usually refers to behaviour that is fitness sacrificing for the helping individual and fitness enhancing for the recipient of the helping behaviour. In philosophy and everyday usage, it refers roughly to behaviour performed with an unselfish concern for others.

means that their history matters. In the following sections I focus on four processes or paths[19] that are generally identified as having played a part in the evolution of cooperation and helping behaviours in animals. These are: kin selection, mutualism, direct reciprocity, and indirect forms of reciprocity.

2.2.1 Inclusive fitness

Inclusive fitness, also sometimes known as 'kin selection', is perhaps the most prevalent form of helping within species of animals that live socially. The basic idea of kin selection is that an organism can increase the frequency of its genes in a population by helping its kin. An organism shares some proportion of its genes with its kin, so one way to improve the reproductive success of its own genes is to improve the reproductive chances of its kin by helping them. The helping behaviours incur some fitness cost to the organism, but as long as this cost is outweighed by the fitness benefits to the helping organism's kin then the net result is evolutionarily advantageous. This relationship was formalized by William Hamilton in his 1964 article[20] and has become known as *Hamilton's rule*. This rule simply states that helping behaviour can occur between kin when the following inequality is satisfied:

$RB - C > 0$

Where R is the genetic relatedness of the helper to the recipient, B is the reproductive benefit gained by the recipient, and C is the cost to the individual that is helping. The outcome of this rule is that we should expect that the frequency and 'value' of the helping behaviours will be proportional to the relatedness of two individuals. Thus, in sexually reproducing species such as our own (with one chromosome of each pair from each parent), the genetic relatedness of siblings is approximately 0.5, parents to offspring is also 0.5, grand-children are 0.25, and cousins 0.125 and we should expect that altruistic behaviour should be more frequent and more substantial between those individuals with

[19] Lee Dugatkin's term for these processes in his 'The evolution of cooperation: Four paths to the evolution and maintenance of cooperative behaviour'.
[20] W. D. Hamilton, 'The genetical evolution of social behaviour'.

higher coefficients of relatedness. This prediction is borne out in the fact that fitness sacrificing helping behaviour in such species is most prevalent between parents to offspring and between siblings and close kin, with the closest genetic relationships engendering the most significant helping behaviour.[21]

Kin selection provides a well understood and well confirmed theory of why many helping behaviours directed at kin are present in a diverse range of species. In what way might this contribute to the explanation of morality (remembering that here we are only interested in the first steps towards, or elements of morality, which can help account for its emergence)? Firstly, familial relations form a significant part of the moral domain. Certainly, a large part of our moral behaviour involves relations with non-kin, yet at least some of our moral obligations are to family and some of these are *unique* to family.

There are however, other more concrete ways in which kin selection contributes to the 'building blocks' of morality, including towards *non-kin*, through the proximate mechanisms it helped construct. The proximate mechanisms of kin selection are the particular systems that allow organisms to recognize kin and regulate helping towards them. Firstly, the system that recognizes kin may not end up 'targeting' only kin. For example, many birds 'imprint' on and thus take to be their parent, the first suitable moving object they see. The range of suitable moving objects turns out to be quite wide, and there are many documented cases of birds imprinting on humans, other animals, and even moving inanimate objects. While some kind of imprinting may or may not be part of the process of kin-recognition in humans, it serves to illustrate that mechanisms for targeting kin can be far from perfect in novel circumstances while being sufficiently effective in their usual environments. In humans, there appears to be a strong propensity to treat those raised with us from a very young age as siblings. This mechanism (like imprinting) may have been highly effective at picking out genetically related siblings

[21] *Ibid*., also see Austin Hughes, *Evolution and human kinship*, for a full treatment of kin selection in human and non-human animals.

in our evolutionary past where populations were structured in small family groups. There is evidence that such a propensity exists in the form of what is known as the *Westermarck effect*. This is when individuals who are raised together from an early age become desensitized to sexual attraction to one another. It has been hypothesized that this is the result of an incest avoidance mechanism and thus requires kin recognition. Evidence for this phenomenon can be seen in Israeli Kibbutzim (with extremely low rates of intermarriage between children raised communally) and in arranged marriages in China (where young girls who are to marry a young male member of a family are adopted and treated as daughters, with the common outcome of failed marriages due to the absence of 'romantic' interest). So, if kin targeting systems are based on a heuristic, such as 'treat as kin those you are raised with from an early age', then it is possible that behaviour that is the result of an adaptation for helping kin may end up being targeted at non-kin in novel environments. As Richard Joyce writes, because "humans now live in societies in which we interact with far more conspecifics than natural selection ever dreamed of (including 'virtual interactions' supplied by TV, newspapers, and so forth), then one would expect to observe ceteris paribus, a great deal of helping behaviour towards non-kin, despite the fact that kin selection is the only explanatory process in play."[22]

The third way in which kin selection may have contributed to morality is through the construction of motivationally powerful proximal mechanisms that regulate helping behaviour. Since natural selection operates on variation of what is already present, it is unsurprising to find old structures that originally developed due to one kind of selective pressure being recruited for other purposes or uses. One non-moral example of such a change of uses for a proximal mechanism is oxytocin, a hormone in mammals that originally regulated maternal nurturing behavior towards offspring, but was later modified to also play a role in regulating pair-bonding behaviour.[23] So, while the proximal mechanisms for helping behaviour may have developed to promote cooperation with kin, there is reason to think

[22] Richard Joyce, *The evolution of morality*, p. 23.
[23] Example from *ibid*., p. 22. See also, Allman, J. *Evolving brains*, p. 97 and 199.

that these mechanisms could have been put to use in subsequent mechanisms for regulating helping behaviour towards non-kin.

2.2.2 Mutualism

By-product mutualism (hereafter simply mutualism) is perhaps the most straightforward process that gives rise to helping behaviours. This is the easiest form of cooperation to explain evolutionarily because it describes a situation where fitness is enhanced for both participants immediately, and the fitness benefits outweigh the cost of participating for both organisms at the same time. An illustration of this kind of cooperation is cooperative hunting. By hunting together, a pair of individuals may be able to capture prey that they would be unable to capture alone. In mutualism there is little opportunity to 'defect' or 'free-ride' in an attempt to attain the benefits of the cooperation without paying the costs: all participants must contribute for the benefits to be available at all and the benefits of cheating are less than the reward available by cooperating. Another important feature of mutualism is that it does not require a stable or ongoing relationship between participants: because participants benefit immediately, it is in their interests to cooperate in such circumstances (where interests means what is fitness enhancing) regardless of whether they participate in future mutualistic behaviour or never interact again.

While the process of mutualism may be conceptually simple to understand as clearly favouring cooperation, it does not follow that the type of behavioural strategies or the proximal mechanisms underlying such strategies are themselves correspondingly simple. Mutualism requires coordinated action on the part of participating organisms, and in cognitively capable species, is likely to produce psychologies that are on the lookout for the possibility of cooperative exchanges. Further, the proximal mechanisms used for mutualism may be modifications of proximal mechanisms for helping kin (a possible first step towards helping directed 'purposely' at *non-kin*), and may themselves have

been modified by other processes for further use, such as in kinds of reciprocity or helping behaviour selected for by group selection (more on this in §2.2.4 and §2.3).

2.2.3 Cooperation

The idea of direct reciprocity is that sometimes it can be in the interests of an organism to behave in ways that advance another organism's fitness, if doing so will result in the other organism returning the favour at some point in the future. This can be for a number of reasons: the value of the help can exceed the cost incurred by the helper (meaning that overall the costs to each participant are outweighed by the benefits), alternatively the benefits may be impossible to attain without such a 'turn taking' practice occurring, or it may be that each organism's contribution is of a specialized kind which provides benefits that would be unattainable without the cooperation. Such a system has enormous potential for enhancing fitness, but it is also one that is difficult to establish and maintain. The difficulty is that reciprocity is vulnerable to exploitation by free-riders: organisms that accept help from others (they get the benefits) without themselves helping in return (they do not pay the costs).

This difficulty is often illustrated using the prisoner's dilemma, a model from game theory. The original thought experiment goes something like the following. Imagine you and an accomplice have been arrested and are charged with committing a crime (whether either of you actually did commit a crime is of no consequence). You are separated from your accomplice, and given the following options:

- If you confess against your accomplice, and your accomplice remains silent, you will go free and your accomplice will be sentenced to 10 years in prison.
- If you confess against your accomplice, and your accomplice confesses against you, you will both be sentenced to 5 years in prison.
- If you remain silent, and your accomplice remains silent, you will both be sentenced to 2 years in prison.

- If you remain silent, and your accomplice confesses against you, you will be sentenced to 10 years in prison, and your accomplice will go free.[24]

These options can be summarized in a matrix to make the situation's structure clearer:

		Accomplice (A)	
		Confess	Remain Silent
You (Y)	Confess	Y = 5 years A = 5 years	Y = 0 years A = 10 years
	Remain Silent	Y = 10 years A = 0 years	Y = 2 years A = 2 years

Figure 1: Prisoner's dilemma

In a one-off prisoner's dilemma, since you do not know what your accomplice will do, the best option is to confess. We can see this by considering the payoffs of your options when your accomplice confesses and remains silent. If he confesses, you have two options. You could confess and you will get 5 years, or you could remain silent and get 10 years. So, if he confesses you should also confess. If he remains silent, again you have two options. You could confess, in which case you will go free, or you could remain silent, in which case you will get 2 years. So, if he remains silent, you should still confess. So, for either option that your accomplice chooses, your prison term will be minimized by confessing.

The structure of the prisoner's dilemma can be summarized by ordering the payoffs from most individually desirable to least.[25]

T > R > P > S

[24] Description based on that given in James Rachels, *The Elements of Moral Philosophy*, p. 148.
[25] Description of ordinal payoffs from Steven Kuhn, 'Prisoner dilemma'.

Here T stands for 'Temptation', the payoff a participant receives when they confess and their accomplice remains silent. R is the 'Reward' that both players receive if they both remain silent. P is the 'Punishment' a participant receives when they confess and so does their accomplice, and S is the 'Sucker's' payoff that a participant receives if they are the only one to remain silent. Any such situation that has the above payoff ordering can be considered to be a prisoner's dilemma. Such situations can be used to model potential cooperative situations in biology if we assume that natural selection will produce organisms that employ strategies to maximise their fitness in much the same way as one of game theory's 'rational agents' will make choices that maximise their own interests. If this is the case, then we can expect that organisms that face a situation where fitness can be maximised by free-riding or defecting will do so. This is why, as Kim Sterelny concludes, "for most animal species, the temptation to defect subverts cooperation."[26]

How then, if cheating is always the best option in a prisoner's dilemma, can cooperation result from such situations? The answer is that such opportunities for cooperation arise not once, but repeatedly. Thus, if an individual with an appropriate strategy can interact on enough occasions with another individual with a similar cooperative strategy, higher long term payoffs are available than in a single instance of a prisoners dilemma. This is because the two options that will be available consistently in an iterated prisoners dilemma are R (both cooperating) and P (both defecting). Any organism that settles for S – the suckers payoff (cooperating while their partner defects) is not likely to last very long; being defected on consistently is not a viable strategy. Thus, since R > P, strategies that employ reciprocal cooperation will result in more evolutionary advantageous outcomes than strategies of reciprocal defection. Those organisms that consistently receive reward payoffs will be fitter than those that receive the punishment outcome.

There is a large body of work on strategies that achieve this outcome as evolutionary solutions to the prisoner's dilemma. The standard starting point for discussion of strategies comes from Robert

[26] Kim Sterelny, *Thought in a hostile world: The evolution of human cognition*, p. 124.

Axelrod's computer simulations of various strategies competing in iterated prisoner's dilemma tournaments.[27] In Axelrod's experiments the strategy of 'tit-for-tat' was the most successful in a range of populations made up of various strategies. Tit-for-tat is a simple algorithm that starts by cooperating and each round follows what its opponent does in the previous round. This strategy reaps the benefits of cooperation (it will cooperate while its partner does – the R payoff), does not allow itself to be exploited (it defects as soon as its partner attempts to do so: resulting in T, and avoiding the sucker's payoff S), and it is forgiving (if a defecting partner starts cooperating again, it will also do so). The outcome of this large (and still expanding) body of work on algorithmic strategies for cooperation is that there are a number of effective solutions to the iterated prisoner's dilemma that consistently favour mutual cooperation over mutual defection. The optimal strategy in any particular 'tournament-like' situation depends on a number of things including what strategies the other individuals in the population use, whether there is any 'noise' introduced into interactions, or any other assumptions that are added to the model to make it more representative of 'real' situations.

The adaptive advantage of cooperation is immense but despite the success of theoretical models, direct reciprocity does not appear to be a widespread process in nature (excluding perhaps some of the higher primates). Marc Hauser provides a relatively comprehensive overview of the empirical evidence for direct reciprocity[28] including studies of vampire bats, blue jays, capuchin monkeys, and cotton-top tamarins, guppies during predator inspection, cooperative territorial defense in lions, grooming among impala and a number of nonhuman primates, coalitions among male baboons and among dolphins for access to females, and food-sharing among chimpanzees. He concludes that either animals do not reciprocate and apparent cases can be explained by other processes (such as mutualism or misdirected kin selection), or if there is reciprocity, then it is uncommon, unstable, or only generated under artificial conditions. While reciprocal behaviour might be an adaptive strategy,

[27] Robert Axelrod, William D. Hamilton, 'The evolution of cooperation'.
[28] See Marc Hauser, *Moral minds: How nature designed our universal sense of right and wrong*, pp. 380-392.

the conditions necessary for it to get off the ground are generally not met. In the very limited cases where helping behaviours such as grooming are exchanged, the reciprocation involves only one single commodity, in a limited and particular context, and the time between helping behaviour and reciprocation is usually very short.[29] So, it seems that direct reciprocity cannot account for the generation of much helping behaviour. While recent empirical work points towards many of the purported examples of reciprocity being cases of the other processes discussed in this chapter, as we shall see in the next section, the theoretical models of reciprocity are still important. While most animals only display at the most, specific kinds of reciprocity with very short time spans between reciprocation, reciprocity among humans is the opposite. It displays a high level of generality and abstractness; one thing may be traded for another of a different kind, and reciprocal relationships are maintained that have long periods of time between acts of reciprocation.

To summarize, reciprocity, if it contributes to the explanation of morality, does so in one of two ways. Firstly, it may have had a role early on, in producing mechanisms to motivate organisms to make basic exchanges of items such as food or behaviours such as grooming – although the evidence for such propensities is limited. Secondly, these motivational structures may be available later on for modification by subsequent processes, which working alongside other cognitive capacities, produce much more abstract and general forms of reciprocal exchange. The next section focuses on this possibility.

2.2.4 Indirect reciprocity

Indirect reciprocity, as its name suggests, introduces an intermediate step into reciprocal interactions. The intermediate step is the exchange of a form of currency for social interaction: reputation. In a prisoner's dilemma, the direct reciprocity solution depends on responding appropriately to the previous interaction. If one player cooperates in one round, it shows they are willing to cooperate in

[29] *Ibid.*, p. 391.

the next and as long their cooperation is met with cooperation from the other player a reciprocal relationship can be maintained. There are reasons to think however, that the prisoner's dilemma is too simplistic to be an adequate model for this kind of interaction in many of the more cognitively developed animals and in humans. The main reasons for this are as follows. Firstly, individuals do not live in an epistemic vacuum; often they will have the opportunity to observe how others interact before they themselves interact with them. They can therefore gain information about what kind of strategy others are using, and what to expect in interactions with that individual. Secondly, individuals within a population of possible reciprocating partners are not (despite the model's name) prisoners: they are not locked into an endless cycle of mutual defections with one partner that will not cooperate. Individuals may cease unproductive interaction with defectors, often while incurring very few costs at all to themselves. This alters the payoff structure considerably: defecting may result in the individual who was cheated withdrawing all possible future interactions with that particular non-cooperator. Thirdly, with reputation, the capacity to punish non-cooperators is much higher. Not only can you withdraw the future opportunity to interact, you can punish even more severely by discouraging other individuals from interacting with the defector.

Once helping behaviours can be traded, reciprocal relationships can give rise to specialisation which makes reciprocity an even stronger strategy. Reciprocity and specialisation can work in conjunction to 'push' each other to greater levels. Such reciprocal behaviour is a non-zero-sum game – if some behaviour costs each organism very little, but helps another significantly, then the benefits to participants can greatly exceed the costs paid by each. The cognitive requirements for this kind of indirect reciprocity however are not insignificant. Individuals need to be able to recognize each other, remember what was traded, and be able to estimate the benefits and costs of such a trade. They need to be able to tell when help is given intentionally, or if it was simply a by-product of some other action or even an accident. Making use of reputations for deciding whether or not to interact with an individual involves being aware of the standing of others, and being able to effectively read and transmit signals about others' statuses. One of the key features that make cooperation through

indirect reciprocity more tenable is the possibility of *cheap* enforcement through the reputation. Enforcement significantly alters the structure of the payoffs from that of a traditional prisoner's dilemma, which with numbers representing fitness payoffs looks something like the following in matrix form:

		Individual A: Defect	Individual A: Cooperate
Individual B	Defect	A = 2 B = 2	A = 0 B = 10
	Cooperate	A = 10 B = 0	A = 5 B = 5

Figure 2: Prisoner's dilemma payoffs without punishment

To a structure where the payoff for defection is made undesirable due to punishment, similar to this:

		Individual A: Defect	Individual A: Cooperate
Individual B	Defect	A = -2 B = -2	A = 0 B = -2
	Cooperate	A = -2 B = 0	A = 5 B = 5

Figure 3: Prisoner's dilemma payoffs with punishment

Clearly in a situation such as this where any defection is punished, the best option is to cooperate. Here the prisoner's dilemma payoff ordering, T>R>P>S, becomes R>S>P>T. Thus, punishment of defection, if it can be established, is an effective way to suppress the temptation to free-ride in cooperative situations. Reputations based on how one interacts with other individuals, and social pressure to conform to cooperative behaviours is one possible way that punishment can be implemented without high costs to those individuals participating in the punishment (this also diffuses somewhat the problem of 'second-order' defection – free-riding in cooperative situations of enforcement). So, the prisoner's dilemma is a useful tool in modelling potential cooperative situations of a less mechanical or algorithmic kind. Due to its cognitive requirements, indirect reciprocity is unlikely to be present in less cognitively advanced organisms, but in conjunction with the processes in the next sections, it is plausible that it played a significant role in the evolution of more advanced kinds of cooperation.

2.3 Evolved normative guidance?

The previous sections discuss the most important processes involved in the development of helping behaviours in the animal world, and how increasing cognitive complexity and plasticity in response to potential cooperative situations could result in more advanced kinds of cooperation. For the most part, biology provides an empirically backed, coherent, and explanatory account of the beginnings and adaptive advantage of cooperation and social behaviour in animals. Clearly however, this is just part of the origin of human morality; there is still a large gap in both magnitude and kind between human social interaction and the social interactions of the rest of the animal world. There is an increasing amount of theoretical work being done on the processes that were involved in getting from cooperative and social forms of living to the ultra-sociality and advanced normative guidance of

humans.[30] It has been theorized that there are a number of processes or elements that were involved in this transition. Morality and normative guidance require language, so one of the key elements that developed was the emergence of language capabilities in the human lineage. Another likely key element was the evolution of uniquely improved forms of social learning which resulted in humans having a pervasive and highly variable culture. It has been hypothesized that this allowed group selection based on cultural variation to take off. In the next sections I discuss first the evolution of language and its contribution to morality and then culture and group selection and their possible contributions.

2.3.1 Morality and the evolution of language

In higher animals and primates, the social problems solved by the processes of kin selection, reciprocity, mutualism and so on, seem to be analogous to the 'subject matter' of moral judgments, and there are good reasons to think that parts of homologous motivational structures are in place in such animals.[31] Some of the most obvious features that are missing from these kinds of cooperation are the cognitive, conceptual, and linguistic elements, and also the authority, and apparent normative force that human morality involves. Cooperation in higher animals and primates appears to be mediated mostly by pro-social urges, but as the discussion of direct reciprocity indicated, the potency of these pro-social feelings is limited: they are vulnerable to breakdown in situations where social defection would benefit an individual. Philip Kitcher provides a good illustration of this, summarizing a case from Frans De Waal's *Chimpanzee Politics* where cooperation is undermined by social defection for self-interested goals:

[30] See William Fitzpatrick 'Morality and evolutionary biology', specifically section 2.3 'Explaining the origins of morality: From psychological altruism to the evolution of normative guidance'.

[31] Frans de Waal provides an argument for this in the form of a principle of *evolutionary parsimony*. He argues that the best explanation of the apparent emotions we see in social interactions in primates is that they really are there: we should not posit new *ad hoc* motivational structures when we already have good candidates for the job. See Frans de Waal, *Primates and philosophers: How morality evolved*, pp. 61-62.

> Researchers spent the daylight hours observing the behaviour of a colony of chimpanzees in Arnhem, Holland. They duly recorded patterns of association, alliances that enabled animals to obtain outcomes they wanted. For some years, two males had supported one another in this way until the two, in concert, dethroned the male who had previously been dominant. At that point, one of the males forsook his old coalition-partner (friend?), pursuing a strategy apparently aimed at monopolizing the females of the colony. This action precipitated a series of intense conflicts, with swiftly changing alliances and profound social instability. In the end, the male who forsook his old alliance was savagely attacked by the former dominant male and the forsaken friend, and the attack proved fatal.[32]

This scenario is not atypical for non-human primate societies: primate social life involves participation in coalitions and cooperation of various kinds, and these relations are frequently disrupted by defection. As Kitcher puts it there is a "delicate interplay of opposing forces – the altruistic dispositions drawing animals to act together and the selfish disruptions threatening to decompose the social group."[33] This disruption results regularly in a breakdown of social cohesion requiring time-consuming repair: usually in the form of mutual grooming. Grooming takes up a considerable amount of time, far more than is necessary or useful for the removal of parasites in primates such as chimpanzees or bonobos. One implication of this is that the amount of time spent cooperating is limited by the amount of time available for social repair: fighting individuals do not make good cooperators. Another implication is that social group *size* is limited. Since grooming is a one-to-one activity, the number of individuals that can groom each other is heavily restricted.

At most non-human primates have about 20% of their 'time-budget' to spend on grooming.[34] At some point however our ancestors circumvented this limitation. One way we can tell that this happened is

[32] Philip Kitcher, '*Biology and ethics'*, p. 171.
[33] *Ibid*., p. 171.
[34] Dunbar, R. I. M., 'Coevolution of neocortical size, group size and language in humans'. This is an empirical finding; when required to spend more time on grooming, primate groups will cease to function and will disperse. Dunbar explains: "Given that primate groups are held together by social grooming, time budget constraints on group size become an important consideration. Even if a species has the cognitive capacity to manage all the relationships involved in large groups, there may be circumstances under which the animals simply do not have the time available to devote to servicing those relationships through social grooming. Relationships that are not serviced in this way will cease to function effectively; as a result, the group will tend to disperse and the population will settle at a new lower equilibrium group size", p. 699.

by looking at brain sizes (more specifically the ratio of neo-cortex volume to the total volume of the brain). In surviving species of primate there is a strong correlation between this neo-cortex ratio and group size: the larger the neo-cortex (proportionally) the larger the group size. As group size increases, the time required for maintenance of social relationships increases. Based on the correlation between group sizes and neo-cortex ratios, estimates of the duration various hominid ancestors would have had to spend on grooming every other member of the group have been made. For *Australopithecines*, estimates place group size at roughly 67, requiring 18% of their time-budgets for grooming; *Homo Habilis* in groups of about 82, requiring 23% of their time; *Homo erectus* at 111 in a group and 31% of their time; and *Homo Neanderthalensis* at 143 and a large 41% of their time.[35] So from the early *Homo* (*habilis and rudolfensis)* onwards, the time required for grooming was more than what was likely to be available.[36]

The outcome of these considerations is that somewhere along the line, hominids must have hit upon a new and much more efficient substitute for social exchange: language. If large group size was sufficiently adaptively advantageous[37] then there would be increasing pressure for more efficient forms of exchanging information about relationships and conduct of group members. Language is a powerful tool for social interaction for a number of reasons. The reputation component of indirect reciprocity immediately becomes more accessible if one can *talk* with others about the trustworthiness of prospective interactants. Compared to individually trying to observe how every other individual acts, conversations about others' conduct are an information goldmine. The amount of information gathered and social 'networking' that can be accomplished in 20% of one's time far outweighs methods requiring direct interaction between each individual. Additionally language is an important instrument for allowing reputation to be used for punishment and commendation. Information can be directly conveyed as to what kinds of treatment people deserve based on what

[35] Leslie C. Aiello, R. I. M. Dunbar, 'Neocortex size, group size, and the evolution of language', pp. 188-189.
[36] *Ibid*.
[37] There are a number of hypotheses as to why this might be so, the most obvious being larger numbers were advantageous in intergroup conflict. See Richard Alexander, *The biology of moral systems*, pp. 79-81.

actions they have made. Also, reciprocal exchanges that involve anything beyond the most basic kinds of 'trades' create pressure for more efficient forms of information transfer. Leda Cosmides and John Tooby illustrate this nicely in the following:

> If I want to exchange an axe for something, how do I indicate what I want? Let's say that I point to the pear you are holding in your hand. What am I referring to by pointing at the pear? Do I want that particular pear? Any pear at all? Five bushels of pears? A fruit of some kind, not necessarily a pear? To be led to the site where you found such good pears?[38]

Thus, language use would be a highly valuable tool in making reciprocal exchanges and may have been necessary for any trades beyond a certain level of complexity.

So, the hypothesis is that language functioned as a more efficient substitute for peacemaking and social bond maintenance activities such as grooming, and additionally it enabled a huge increase in the exchange of information, especially about the social conduct of group members. A large part of the adaptive value of such communication is that people with poor reputations can be avoided, shunned, or punished, and those with good reputations interacted with and commended to others as good co-operators. The kind of information conveyed in conversations, was for the most part, not simply descriptions of behaviour: the purpose of the communication was to criticize and commend other's behaviour. As Richard Joyce amusingly writes "Effective (juicy) gossip involves more than mere *descriptions* of who did what to whom; it embodies praising and condemnatory language – perhaps along the lines of 'Ogg never repaid Gak for that axe: the *scoundrel*!' (to choose a rather quaint translation) or 'Klug always repays a favor: He's a great guy!'"[39] So, right from the outset, it is plausible that language used in social contexts may have conveyed both information (not repaying an axe, or being someone who reliably repays favours) and also evaluative content (he's a scoundrel, a great guy).

[38] Leda Cosmides, John Tooby, 'Evolutionary psychology and the generation of culture, part II: Case study: A computational theory of social exchange", p. 64.
[39] Richard Joyce, *The evolution of morality*, p. 91.

The evolution of language was clearly an important element in enabling the possibility of normative guidance – a capacity to establish rules and have those rules impact on the wishes, plans and intentions of all the members of a group in a way that permitted high levels of cooperation and collective action.[40] Without language, morality as it exists in humans would not be possible: morality and normative guidance require *evaluative* concepts, and further, these must be communicable.[41] For example, to formulate the idea that Ogg is a *scoundrel* requires a kind of conceptual complexity that is only available to language users and the communication of such information is only feasible for language users (imagine trying to communicate that you saw someone take more than their fair share of food without being able to use any of our linguistic capacities). Once established, normative guidance enabled the group size limits to be transcended and induced more stability into what was previously a somewhat fragile and turbulent early hominid social climate. This transition allowed the change from pre-moral primates and hominids that simply accepted or did not accept particular behaviours and reacted in either hostile or friendly ways, to organisms that had an awareness of whether particular behaviours broke certain norms, and that particular behaviours will not only provoke hostile or friendly reactions from other group members, but also *merit* or *deserve* particular reactions.

Language is an extremely important and integral element of morality and social living in humans. Language made it possible to easily express, memorize, and exchange information, evaluations, and concepts. It allowed huge increases in productivity and the number of problems solvable cooperatively. While the evolution of language was a necessary component for the evolution of morality, the evidence regarding *when* language evolved is relatively inconclusive even relative to other processes such as the development of culture. The fossil evidence is ambiguous and scarce, with some claiming that there are identifiable brain structures associated with language present in hominids from two million years ago, while others argue that the soft anatomy of the vocal tract

[40] Philip Kitcher, *Biology and ethics*, p. 172.
[41] For a discussion of this, see Richard Joyce, *The Evolution of morality*, pp 80-85.

indicates that even very recent hominids, perhaps only 50,000 years ago still may have only had extremely limited speech.[42] Additionally, it is not easy to find other empirical evidence that is relevant to finding out how or when the transition from non-linguistic to linguistic communities took place. For these reasons, while the evolution of language seems to be highly relevant for explaining how morality developed, accounts of its role are still highly theoretical and somewhat speculative.

2.3.2 Group selection and culture

Multilevel selection has been the subject of considerable debate in biology in the last 40 years, during which there has been little consensus as to its status. There is however something of a trend towards its acceptance, especially in relation to its role in explaining the evolution of new levels of hierarchical organisation in the biological world. Examples of these transitions where lower-level entities aggregate to create new higher-level entities include the transitions from individual genetic molecules to chromosomes, prokaryotes to eukaryotes, single-celled organisms to multi-celled ones, and in group selection (the transition that this section focuses on) from individual organisms to cooperative groups of many such organisms.[43] A striking feature of these transitions is that while they involve selection operating at both levels simultaneously, they ultimately result in a high degree of functional integration (i.e. cooperation) and suppression of competition between the lower-level entities.[44]

While group selection is somewhat contentious, some like Kim Sterelny have argued that it played a pivotal role in the extraordinary levels of cooperation found in humans.[45] The thought here is that group selection became important once a certain level of social capability was present, and once it did take off, it allowed for even more comprehensive and pervasive kinds of cooperation to develop. Sterelny puts forward three conditions that must be met if group selection is to be a powerful

[42] Peter Richerson, Robert Boyd, *Not by genes alone*, p. 144.
[43] Examples from Samir Okasha, *Recent work on the levels of selection problem*, p. 350.
[44] *Ibid*., p. 350.
[45] See Kim Sterelny, *Thought in a hostile world: The evolution of human cognition*, pp. 123-145.

evolutionary process.[46] Firstly, there must be variation in the groups' levels of cooperativeness (i.e. something that allows them to be differentially successful). Secondly, cooperative individuals must have a tendency to form groups with other cooperative individuals. And thirdly, the fitness advantage bestowed by being members of cooperative groups must outweigh the fitness advantage of selfish individuals over cooperative individuals in mixed groups. There are generally two ways this third condition can be met: either the fitness benefits to those that free-ride over those that do not are relatively small compared to the fitness advantages at the group level, or there are barriers which stop free-riding or make it a costly choice.

There are a number of reasons to think that humans and our predecessors satisfy these conditions. Social learning and primitive forms of culture can provide large variation in the fitness levels of groups. Indeed the variation between different groups is one of the notable features of humans: neighbouring groups of modern humans can have radically different cultures including different languages, dress, social structures, technological advancements, religious beliefs, foods, and customs. Having a tendency to form groups and cooperate with other individuals who are part of the social group is a well-established tendency in human primate ancestors. Finally, human cognitive abilities can also achieve both the flattening of fitness advantages of free-riders, and impose barriers to free-riding taking place: the monitoring of cheaters and imposing of sanctions on them can significantly alter the cost to benefit ratio so as to make free-riding not worthwhile.[47] These kinds of punishment or enforcement behaviours need not be of the costly or dangerous kind either, as Richard Joyce reminds us "if the punishment is the withdrawal of social esteem, which can be distributed or denied like a magical substance, or exclusion from ongoing beneficial exchanges, then punishment can often be meted out at no cost."[48]

[46] *Ibid*., pp. 125-126.

[47] The details of cooperation for enforcement are complicated – enforcement itself is a cooperation problem and therefore immediately raises the possibility of so called second order defection problems. Kim Sterelny offers a good introduction to these difficulties and the processes that may have overcome such difficulties. See *Thought in a hostile world,* p. 126.

[48] Richard Joyce, *The evolution of morality*, p. 41.

Thus, the advent of our species' unique capacities for culture is thought to have had substantial implications for the development of the uniquely social forms of living in humans. Culture here simply means any "information capable of affecting individuals' behaviour that they acquire from other members of their species through teaching, imitation, and other forms of social transmission."[49] Social learning and imitation, resulting in distinctive cultures, tends to increase the variation between groups, while simultaneously reducing the variation within groups.[50] Significant variation between groups and homogeneity of individuals within groups, can allow for powerful selection at the level of the group to take place. Peter Richerson and Richard Boyd argue that the human capacity for social learning and cumulative culture, via a kind of population-structured selection based on cultural groups, gave rise to a new kind of social world – an environment that drove the selection of novel social instincts in the human lineage. For example, if group selection as a result of the varying cultural practices in groups creates an environment where cooperation is rewarded and non-cooperation is heavily discouraged and punished, then the climate in which gene selection is to take place has changed dramatically: genetic selection in such conditions will be likely to result in psychologies strongly disposed towards cooperative social behaviour. This interaction between the various processes and the resulting pressure towards pro-social behaviours operating on organisms with already functional mechanisms for motivating social interactions resulted in the cooperation explosion that is unique to humans. This kind of theory of simultaneous and interacting cultural and genetic inheritance is sometimes called a "dual inheritance theory."[51]

Richerson and Boyd argue that this kind of co-evolution of culture and human psychology has a deep evolutionary history, and as a result we have deeply entrenched "tribal" instincts to go along with the

[49] Definition from Peter Richerson, Richard Boyd, *Not by genes alone*, p. 5.
[50] Kim Sterelny, *Thought in a hostile world*, p. 127.
[51] Sterelny describes 'dual inheritance' as follows: "Children resemble their parents because of the flow of genes from parents to children. But children also resemble their parents because there is an extensive and accurate flow of information from parent to child", Kim Sterelny, 'Review: *Genes, memes and human history*', p. 250.

pro-social instincts that evolved earlier in our phylogeny.[52] By "tribal" instincts they mean a suite of new social instincts suited to life in groups,

> …including a psychology which "expects" life to be structured by moral norms and is designed to learn and internalize such norms; new emotions, such as shame and guilt, which increase the chance the norms are followed; and a psychology which "expects" the social world to be divided into symbolically marked groups.[53]

So, the role of group selection is generally thought to have been influential later on than the processes leading to cooperation discussed in previous sections. It required cognitive capacities that were most likely the result of these earlier processes and depended upon the variation between groups (and uniformity within groups) which was due to cultural evolution. This altered environment produced strong selective pressures for a "tribal" social psychology. Richerson and Boyd think that without including cultural causes, the explanation for human's unique sociality is incomplete.

The theory they present is plausible, and highlights the possibility that culture and social learning had a much more direct influence on genetic evolution than has previously been attributed to it. While the anthropological data they provide for cultural evolution as a process taking place in both historical and contemporary times is compelling, the support for the further claim that culture changed the selective environment in ways that resulted in 'tribal' social instincts evolving, is lacking. Data on hunter gatherers from contemporary or recent historical sources is unable to provide support for the operation of the hypothesized processes in prehistoric times. As Stephen Shennan has commented, it still remains "unclear whether we can apply Richerson and Boyd's models in a reasonably empirically constrained way to explain the deep human history that archaeologists try to study."[54] Additionally, Richerson and Boyd readily admit that "paleoanthropologists have no idea when human language evolved"[55] which is undoubtedly important to the development of social systems based on norms and

[52] Peter Richerson, Richard Boyd, *Not by genes alone*, p. 196.
[53] *Ibid*., p. 214.
[54] Stephen Shennan, 'Not by genes alone: How culture transformed human evolution, by Peter J. Richerson and Robert Boyd', p. 297.
[55] Peter Richerson, Robert Boyd, *Not by genes alone*, p. 144.

more generally culture of any complex kind. Richerson and Boyd recognize that their account of human ultra-sociality is still highly theoretical, and many of the details are yet to be filled in, or are likely to change.[56] Despite this, they think that any modifications or improvements on their theory will retain many of elements that give it its unique form: being an evolutionary explanation that synthesizes the genetic and cultural causes to explain the special kinds of advanced sociality found in humans.

2.3.3 From sociality to morality – epistemic difficulties

The plausibility of evolutionary theories of sociality developing into something more like normative guidance and morality can be readily established. However, the plausibility of such accounts does not imply we know the actual details of this transition with any certainty or can verify the truth of those accounts. Empirical constraints are both difficult to identify and access, given the historical nature of the transition and intangibility of the phenomena in question. Accordingly, it is important to understand the limits to the certainty that we can place on evolutionary explanations of morality and more generally what we can know of how human cognition evolved.

Richard Lewontin has forcefully argued that the details of the best accounts of some areas of our evolutionary history are necessarily speculative and that adequate confirmation of them may never be found.[57] This is not to deny cognition did emerge during our evolution, as Lewontin agrees that "…it must be, that human cognition, like every other characteristic of the human species, has arisen during the continuous course of human evolution." [58] Nevertheless, providing solid epistemic foundations for evolutionary explanations of human cognition is an exercise fraught with difficulty that may mean we cannot verify the details or even truth of the explanations.

[56] *Ibid*., p. 235.
[57] Richard Lewontin, 'The evolution of cognition: Questions we will never answer'.
[58] *Ibid*., p. 108.

2.3.4 Adaptationism

Stephen J. Gould and Lewontin in 'The spandrels of San Marco and the Panglossian paradigm' highlight what they perceive as a range of systematic errors within what they call the 'adaptationist paradigm'. By adaptationist paradigm they mean an observed tendency of evolutionary theorising to treat adaptation by natural selection as a near omnipotent force; that every existing trait can be explained by and is the result of natural selection. If any constraints are taken into consideration, they are briefly acknowledged but then dismissed or simply ignored thereafter.[59] Lewontin and Gould argue the result of this approach is that evolutionary explanations of phenomena are accepted too uncritically and with insufficient justification or support for the certainty that they are advanced with.

Lewontin and Gould identify a number of what they believe to be common errors in the evolutionary reasoning of the adaptationist programme and identify a number of alternatives to explanation by adaptation and natural selection that they believe are sometimes in operation but are ignored. The faults in evolutionary reasoning they identify include:

- The failure to distinguish current usage of a trait from the potential origin of that trait ("male tyrannosaurs may have used their diminutive front legs to titillate female partners, but this will not explain why they got so small."[60])
- The failure to be critical of evolutionary explanations when they are implausible or poorly reasoned. The fact that an evolutionary explanation can be imagined does not mean it is a good or reasonable one.
- The lack of attempts to evidentially constrain or provide support for adaptive stories as explanations of traits other beyond the plausibility of the story being told.

[59] Stephen J Gould, Richard Lewontin, 'The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme', p. 585.
[60] *Ibid*., p. 581.

They also fault the adaptationist programme for not being willing to consider the variety of alternative explanations to a trait being an adaptation selected for by natural selection. Alternatives include[61]:

- That the trait may not have been an adaptation at all and not be due to selective pressures- it may have some other causal history.
- That the trait may be a by-product of another trait that itself is an adaptation due to natural selection. This is where the reference to the "Spandrels of San Marco" comes from in Gould & Lewontin's article: spandrels are an architectural by-product of arched roofs; they are space between the shoulders of adjoining arches and are necessary if arches are present despite not being the intended or designed feature themselves.
- It is possible that one of selection or adaptation is present but not the other. A trait maybe an adaptation without the explanation of its selection being the real reason it was selected for, or the selective pressures identified may be correct, but the trait itself is not in fact a result of those selective pressures.

Many of these cautionary points are certainly relevant in the case of the evolution of human cognition and behaviour. Indeed, perhaps more so than in many contexts, alternative explanations and causes are likely to be present in the history of human development and there needs to be sufficient evidence to discriminate between those explanations.

2.3.5 Questions we will never answer?

Lewontin goes further in his article 'The Evolution of Cognition: Questions We Will Never Answer'[62] in discussing the difficulties of sufficiently grounding evolutionary explanations of cognition. Lewontin

[61] *Ibid*., pp. 590-593.
[62] Lewontin, 'The evolution of cognition: questions we will never answer', p. 108.

argues that there are three things that must be known to have a defensible explanation of how a trait has evolved via natural selection[63]. The three items are:

a. We must know how a trait varies (The 'Principle of variation')
b. We must know how the trait is heritable (The 'Principle of heredity')
c. We must know how the trait will increase an organism's fitness when it possesses the trait. (The 'Principle of natural selection')

All three of these principles are necessary parts of an explanation of adaptation by natural selection. Without variation, there is nothing to select. If variation is not heritable, then the trait cannot persist or affect the next generation. And if there is no differential reproduction and survival there can be no change in the frequency of the trait in successive generations.

In the case of the evolution of human moral or normative capacities these requirements or principles translate into the following:

a'. That there is variation in the cognitive capacity for normative guidance – it must be demonstrable that some individuals have cognitive capacities for normative guidance that are not shared by other individuals or groups.
b'. That a cognitive capacity for normative guidance that has appeared in some individuals is a heritable trait that will be transmitted to subsequent generations of those individuals who possess it.
c'. That those individuals who have the cognitive capacity for normative guidance will leave more offspring than those lacking it.

It would be fatal to the explanation to be missing any one of these elements of the evolutionary process, but Lewontin argues that we are in fact missing evidence for all of a', b', and c'.

[63] *Ibid*., p. 109.

For a' (the requirement that we must know how a trait varies) the period in which it appears much important cognitive development was supposed to have occurred did not involve visible morphological changes – there are no changes in morphology for the physical fossil record in the most important last 100 000 years, and very little other physical evidence is available to go on. The scarce cultural artefacts and cave drawings might count as one useful signpost of "a cognitive activity of a very advanced nature"[64] but these early forms of evidence are difficult to infer anything reliable from; and single signposts many generations apart are certainly not enough on an individual or group level to provide sufficient evidence to "know how a trait varies" in a way useful for assessing evolutionary change.

Secondly for b' (the requirement that we must know how the trait is heritable) we have very little visibility of evidence showing what or how differential cognitive capacities for normative guidance were passed from one generation to the next and certainly not at a level of detail that would allow us to assess or follow the phylogeny. While suggestive, using comparisons between our closest evolutionary relatives as evidence is also problematic if we do not know whether particular traits are homological (inherited from a common ancestor) or analogical (where traits that are similar evolve because of the similar environmental challenges and selective pressures, but there is no common ancestor). The paradigmatic example of this is linguistic ability in humans and our nearest evolutionary relationships. There appear to be some elements of language that we can teach to chimpanzees and other close primates, but this itself is not evidence that these particular abilities share an evolutionary origin (that they are in fact homologous). It may be that the capacity for such learning evolved in parallel due to similar pressures after we split from a common ancestor – they may simply be analogous capacities. The upshot of this is that what little we might deduce from the differences between ourselves and our nearest ancestors are not sufficient to meet the requirement of b'.

[64] *Ibid*., p. 115.

Finally, for c' (the requirement that we must know how the trait will increase an organism's fitness) there is no evidence available to us showing that individuals with the cognitive capacity for normative guidance had a reproductive advantage within the relevant time period. While it might seem easy for us to imagine how cognitive capacities would result in an adaptive advantage, this alone is not helpful. Firstly our imagination of what might be evolutionarily advantageous may be misleading; as Lewontin rightly reminds us "the view of our individualistic, competitive, and entrepreneurial society that the smart and articulate win power may not apply to our primitive ancestors."[65] But even if we did have an accurate picture of the kinds of cognitive capacities for normative guidance and how they might have made those who possess them more evolutionarily successful (which the accounts in the previous sections suggest we do) we do not have the evidence to confirm this picture or support one potential story over others. There is no obvious way of verifying what made particular individuals or groups who possessed differing cognitive capacities more successful than other individuals and groups who lacked them.

Lewontin's critique of evolutionary explanations of human cognition is a significant challenge for evolutionary accounts of human morality. It suggests that the existing evidence we have will not be sufficient to provide verification for the kinds of theoretical accounts presented in this chapter. Noam Chomsky has also warned against assuming that all questions we can formulate about ourselves will be knowable or verifiable. In 'The mysteries of nature: How deeply hidden?', Chomsky aims to remind modern thinkers[66] that there are limits to what we can know. He specifically mentions the evolutionary origins of cognition as one of these potential 'mysteries' for human understanding.[67] He aims re-popularise the idea that "there is no reason to believe that humans can solve every problem

[65] *Ibid*., p. 129.
[66] Chomsky contends that earlier thinkers including Descartes, Galileo, Newton, Locke, and Hume among others were more cognisant of the limits of human conceivability and what might be known. See Noam Chomsky, 'The mysteries of nature: How deeply hidden?', pp. 174-175.
[67] *Ibid*., p. 199.

they pose or even that they can formulate the right questions; they may simply lack the conceptual tools."[68]

However, in the same discussion Chomsky also warns against the dangers of hastily assigning anything to the category of "unknowable".[69] He argues that problems we identify as 'hard' (including the problem of consciousness often referred to in philosophy as "the hard problem"[70]) are not necessarily more uniquely difficult or unsolvable than other problems we have been able to resolve. The lack of conceivability of a solution at any point in time should not be used as a guide to what is knowable or will be knowable in the future.

There is more than one way in which this warning is applicable. In some cases, a claim of the impossibility of knowing something can be shown to be straight-forwardly incorrect when claims are disproven some years later. This can occur when the phenomena in question becomes understood via new methodology or related discoveries open new theoretical and evidential avenues of inquiry for the original question. An example of this is Auguste Comte's claim in 1835 that we would never know anything about the composition of stars.[71] Comte stated, concerning stars, that "we would never know how to study by any means their chemical composition, or their mineralogical structure".[72] The theories and tools that would allow these measurements had not been conceived of at this point, but by the mid-20th century, after a number of key advancements in various branches of physics and spectroscopy, it was possible to determine the composition, temperature, and structure of stars.

Alternatively, it may be that concepts and theories that an 'unknowable question' depends on are themselves revised or shown to be incorrect which opens up the way for the unknowable concept to be re-examined as a different question. The theory of chemical bonds (or more accurately a lack of a

[68] *Ibid*., p. 185.
[69] *Ibid*., p. 171.
[70] *Ibid*., p. 177.
[71] See J. Hearnshaw, 'Auguste Comte's blunder: An account of the first century of stellar spectroscopy and how it took one hundred years to prove that Comte was wrong!'.
[72] *Ibid*., p. 90.

theory) and the inability to reduce chemistry to known physical phenomena well into the 20th century is Chomsky's example of this. There was a rich field of Chemistry with many experimental and theoretical successes, but no understanding of how it related to or cohered with the rest of physical science- no reduction of chemical to physical theory was viewed as possible. Understanding of the relation turned out to be impossible because the understanding of physics was deficient. The unification between chemistry and physics had to wait until physics had undergone radical changes- only with an updated molecular mechanics based on quantum theories of matter was the reduction possible.[73]

Thus, Chomsky makes two points relevant to the present discussion. Firstly, as per Lewontin, there are things that we may not ever have sufficient evidence for, including explanations of the evolution of human cognition. Second, while there are things we may not know, and things that we may not have evidence for presently, it is difficult to know if this will always be the state of affairs. It is dangerous to assert something is entirely unknowable in the future as the history of science is littered with examples where this has proven to be false.

2.3.6 "How possibly" explanations

Returning to evolutionary accounts of morality, we know that while the theories presented may appear plausible, there is limited evidence to support preferring any given evolutionary story presented over alternative hypotheses. The period of interest in human pre-history, from the first use of tools by *Hominins* approx. 2.6Ma in the early Palaeolithic down to the early Neolithic approx. 10,000 years ago, offers only archaeological evidence.[74] We can build on this with comparisons of anthropological knowledge of human groups who still live as hunter-gathers in group sizes similar to those the fossil record shows. Additionally, comparisons with other primates, our nearest evolutionary

[73] Noam Chomsky, 'The mysteries of nature: How deeply hidden?', p. 187.
[74] See Nicholas Toth, Kathy Schick, 'Overview of Palaeolithic archaeology, p. 1943.

relatives provide insight into the social problems likely faced by groups of similar sizes prior to the Lower Palaeolithic. But these traces and inferences do not provide any certainty and are not sufficient to constrain the evolutionary hypotheses about the development of capacities involved in our moral psychologies.

Given this epistemic situation, Philip Kitcher has suggested that evolutionary explanations that address the development of morality should be conceived of as 'how possibly' explanations.[75] By 'how possibly' explanations he means ones that allow us to explore 'how a sequence of events could have happened' given the constraints of a particular theory, while being conscious that the explanation is ultimately underdetermined (that is, the evidence is not sufficient to establish the explanation as the only one that could fit with the limited evidence we have). This allows explanations to at least establish plausibility, and allows for hypothetical exploration of the consequences were the proposed theory true. Of course, it would be better if we were able to know definitively the actual history of our social and normative capacities, but "…given the temporal remoteness of the events and the limitations of our evidence, modesty is required. In the context of rebutting the skeptical challenge, modesty—settling for "how possibly"— is enough."[76]

There are few points to note about this strategy. Firstly, that this approach is not unique within evolutionary theory to the case of human cognition. There are many areas where evidence is too historically remote or sparse to sufficiently constrain hypotheses, but explanations are still proffered as a best available effort.[77]

Second, while there is not space in the present thesis, it would be an interesting exercise to explore and delineate what the other alternatives to the evolutionary account are, and evaluate the fit of

[75] Philip Kitcher, *The Ethical Project*, pp. 9-13.
[76] *Ibid*., p. 12.
[77] The example Kitcher cites (*Ibid.,* p. 12) is the evolution of the cell – it may be questioned whether we have sufficient epistemic access to be confident that evolution by natural selection was the process involved or if it was, what the details of that explanation are. See also Gijsbert van den Brink, Jeroen de Ridder & René van Woudenberg, 'The epistemic status of evolutionary theory' for more general discussion of where adaptation by natural selection is underdetermined.

these options against what evidence is available, and examine whether they provide the same explanatory or predictive power as the evolutionary account. Building on this exploration, it may be that some of the arguments that attempt to draw conclusions for philosophical ethics from the evolutionary accounts could also apply to non-evolutionary alternatives if they are similar in enough aspects. In some cases, it will simply be sufficient for there to be an account that is 'naturalistic' in that it rules out any elements requiring divine intervention or supernatural influence. In others that particular behaviours or phenomena had the functions they did due to evolution will be relevant. Ultimately, in evaluating whether knowing the exact details of an evolutionary explanation is important and whether a "how possibly" explanation is sufficient will have to be evaluated on a case by case basis. As will be included in the analysis in Chapter 4, the level of detail required of the evolutionary explanation is a consideration that needs to be taken into account as part of assessing any evolutionary account of human sociality and its influence on philosophical ethics.

Given the above, I proceed with the following as a self-consciously "how possible" account. An overall sequence of the hypothetical development of morality might be as follows. Our distant ancestors, the earliest hominids were social animals who lived in groups based mostly on kinship. Later hominids, also our ancestors, evolved language as a response to an increased need to communicate in social situations, and perhaps also developed the pre-cursors to moral emotions and the capacity to make and communicate evaluative judgments. This allowed them to live in cooperative societies that were governed by socially shared rules. As Kitcher puts it, they invented a "*proto-morality*, perhaps little more than some judgments about who belonged with whom and a few crude injunctions about loyalty and revenge".[78] Once this was in place, it allowed other processes, perhaps group selection and indirect reciprocity, to combine with cultural evolution based on variations in culture between groups, resulting in a range of different kinds of systems of rules or norms. These systems were transmitted both genetically (in the form of dispositions to learn language, have certain pro-social emotions,

[78] Philip Kitcher, *Biology and ethics*, p. 173.

preparedness to pick-up evaluative concepts, and so on) and culturally, through oral tradition and inculcation of group wide norms or rules from one generation to the next. In this radically different environment, genetic selection might continue to select for traits that favoured highly social and altruistic tendencies. Cultural evolution, with the differential success of a variety of cultures or "experiments in living" being more successful than others, resulted after many thousands of years in distinct cultural lineages.[79] These cultural lineages, once they reach the limits of our known historical record, become recognizable as including systems of morality.

The history of the later stages of cultural evolution is continuous with the oldest extant oral and written records of religious, moral, and legal systems that we have access to today. Examples of these include: "Early Mesopotamian law codes, versions of myths (the Gilgamesh epic is a prominent example), and the Egyptian *Book of the Dead*"[80] which all contain recognizable moral rules and ideas. While the early evidence of morality is often fragmentary and highly specific to particular situations, the primary function of the rules they contain is to resolve situations of potential conflict. They often concern specific social situations, such as "the causing of miscarriages to the daughters of others, the failure to use an orchard one has rented, [and] the joint maintenance of irrigation systems."[81] This high level of specificity is probably due to the incomplete nature of our record of them, but also due to the fact that they are intended as additions to a widely known, system of social rules that was already a well-established and entrenched part of the cultural inheritance.

In this chapter I have discussed the proposed evolution of morality, via various processes, beginning with the most basic such as kin selection, mutualism, and reciprocity, all of which result in various kinds of biological altruism.[82] Living in social groups allowed the possibility of further processes such as indirect reciprocity, and group selection to result in more extensive kinds of altruism that were mediated by increasingly complex, psychological capacities. The development of language, moral

[79] *Ibid*.
[80] *Ibid*.
[81] *Ibid*.
[82] Altruism in the biological sense of fitness enhancing for others, fitness reducing for the individual helper.

emotions, and evaluative concepts, allowed for group norms to guide behaviour more extensively. Many of the details of this story may be altered in the future as more work is done and theories are modified and developed, and further evidence if it can be identified is taken into account. Undoubtedly there is much more to be said on the issue, but the previous sections should at least make plausible that morality could be in part an evolved capacity, and make pertinent the question of '*if* our psychology so disposed to morality has such evolutionary origins, what are the implications?'

Chapter 3 Implications of the evolution of morality for ethics

In this chapter I look at theorists who have attempted to show that a growing understanding of the origins and background of morality has significant consequences for moral philosophy. First, I look briefly at the attempts by the biologist E. O. Wilson to draw ethical conclusions from the evolutionary origins of human nature and sociality. The term 'Sociobiology' was coined by Wilson to refer to the study of the biological determinants of social behaviour, based on the theory that such behaviour is often genetically transmitted and subject to evolutionary processes. Wilson's *Sociobiology: The New Synthesis* attempted to provide a "systematic study of the biological basis of all social behaviour"[83] and was an impressive synthesis of evolutionary theory and ethology. The majority of *Sociobiology* covered non-human animal behaviour, and only in the last chapter did Wilson turn to discuss the evolutionary explanations of human behaviour. When first published, the comments on human morality and his claims about ethics were highly controversial and open to various kinds of criticisms due to being too fast or incomplete. Wilson followed up in an attempt to clarify many of these ideas with his book *On Human Nature* which attempts to clarify how such sociobiological ideas might apply to human society. In *On Human Nature*, Wilson discusses a wide range of human social phenomena including Religion, War and Aggression, Altruism, Sexuality, Diversity, and the potential merging of social and biological sciences.

It is worth noting that while the term sociobiology is generally not used in discussions about the evolution of human behaviour or sociological phenomena, this is mostly due to it being replaced by 'evolutionary psychology'. The difference between the two is simply that sociobiology was focused on evolutionary explanations of sociality in animals (including humans) whereas evolutionary psychology is focused on only human evolution and the resulting mental adaptations that mediate human behaviour.

[83] E. O. Wilson, *Sociobiology: The new synthesis*, p. 4.

Wilson argues that evolutionary biology can shed light on a number of issues: on the metaphysics of morality, on what he calls the 'problem of altruism', on biological constraints on what we ought to do, and on the 'naturalness' of various human behaviours and consequently their ethical status among other issues. Wilson's arguments are suggestive of a number of different ways in which an understanding of our biological origins may provide novel insights into ethics. While suggestive, I conclude Wilson's arguments are either unsuccessful or underdeveloped, and can be used to highlight a number of common difficulties or complications in drawing ethical conclusions from evolutionary facts. His work shows that useful philosophical conclusions do not simply fall out of the facts about morality's evolutionary genealogy and require argument and engagement with the philosophical literature to make progress.

Secondly, and in more detail, I examine the work of Richard Joyce and Sharon Street who provide sophisticated and in-depth attempts at drawing implications for philosophical ethics from the evolution of morality. Joyce argues that because we have evolved to have certain moral beliefs due to the evolutionary advantage they gave, moral truth may not have played any part in the genealogy of morality and consequently our moral beliefs are in danger of having their justification undermined. In contrast to Wilson, Joyce locates his argument in the wider meta-ethical context and addresses the philosophical difficulties involved in arguing for his position, and his evolutionary argument raises novel considerations for meta-ethical debates about moral scepticism.

In a related topic to Joyce's argument, Sharon Street puts forward a Dilemma for the Moral Realist that aims to challenge them to choose between dropping their commitment to a scientifically defeasible account of the evolution of morality and rejecting their theoretical commitment to moral realism. According to Street's argument, if the Moral Realist wishes to avoid the latter choice, they must argue that somehow the independent moral truth that Realists are committed to was involved in the evolutionary genealogy of morality. No realist has yet met that challenge, and indeed it is hard to see how they would go about doing so. If the Moral Realist wishes to reject that independent moral

truth played a role in the evolution of morality, then they must either accept anti-realism or claim that a coincidence of fantastic proportions has occurred and our evolved moral psychology just happens by chance to be identical to the independent moral reality.

3.1 E. O. Wilson and sociobiology

> The biologist, who is concerned with questions of physiology and evolutionary history, realizes that self-knowledge is constrained and shaped by the emotional control centers in the hypothalamus and the limbic system of the brain. These centers flood our consciousness with all the emotions – hate, love, guilt, fear, and others – that are consulted by ethical philosophers who wish to intuit the standards of good and evil. What we are then compelled to ask, made the hypothalamus and limbic system? They evolved by natural selection. That simple biological statement must be pursued to explain ethics and ethical philosophers, if not epistemology and epistemologists, at all depths.[84]

Due to the strong reaction of many to his work, E. O. Wilson was perhaps one of the most infamous of modern researchers to recognize and comment on the importance of our evolutionary origins for human behaviour. He contended that our increasing knowledge of the origin and evolutionary biology of social behaviour meant that "Scientists and humanists should consider together the possibility that the time has come for ethics to be removed temporarily from the hands of the philosophers and biologicized."[85]

In his writing on this issue there are roughly four distinct claims he makes that indicate the direction this 'biologicization' should take.[86] Firstly, biologicizing ethics means recognizing that moral judgments are the products of biological causes that are, as Wilson puts it, "shaped by the emotional control

[84] E. O. Wilson, *Sociobiology: The new synthesis*, p. 3.
[85] E. O. Wilson, *Sociobiology: The new synthesis*, p. 562.
[86] In E. O. Wilson, 'Man: From sociobiology to sociology', Chapter 27 in *Sociobiology*, and E. O. Wilson, *On human nature.*

centers in the hypothalamus and limbic system of the brain".[87] The consequence of recognizing this fact is that the metaphysics of morality will be demystified: judgments of morality are no longer to be considered "occult truths known through moral intuition."[88] Thus Wilson thinks that applying Sociobiology and evolutionary theory to ethics will shed light on the ontological and metaphysical problems of ethics.

The second problem Wilson thinks evolutionary theory solves is what he calls the 'problem of altruism'. He argues that evolution can explain why humans are altruistic; why it is that we sometimes act in ways that benefit others at a cost to ourselves. Evolutionary theory can explain why humans are such highly social creatures, why we make moral judgments and why doing so would have been adaptive.[89]

Wilson's third claim is that in some sense we must "adjust our ethical judgments to fit the realities"[90] that a sociobiological analysis reveals about ourselves. The idea here is that if a sociobiological analysis shows that we are genetically predisposed to be helpful or considerate to only friends and family say, and not at all altruistic to distant strangers, then because these tendencies or behaviours are genetic predispositions, there is little point in attempting to alter them: our biology *constrains* what we ought to do by constraining what is possible for us to do.

Finally, the fourth direction that Wilson suggests the biologicization of ethics should take is that biology can inform us of whether something is a naturally occurring phenomenon and this information can be used in evaluating its status. For example, Wilson's discussions of sexuality and homosexuality at times appear to endorse claims of the following kind: because these are naturally occurring practices, they should be considered acceptable or unobjectionable, and we would therefore be

[87] E. O. Wilson, *Sociobiology: The new synthesis*, p. 3.
[88] James Rachels, *Created from animals: The moral implications of Darwinism*, p. 77.
[89] E.O. Wilson, *On human nature*, pp. 149-167.
[90] *Ibid*., pp. 141-148.

wrong in attempting to alter these patterns of behaviour.[91] In the following sections, I discuss each of these four claims in turn.

3.1.1 The metaphysics of morality

By recognizing the biological basis of morality, Wilson thinks that progress in metaethics concerning the metaphysics of morality can be made.[92] This does not appear to be an unreasonable claim; an empirical approach seems like it would enable us to identify the kinds of entities that are involved in morality and assess which of those are scientifically more or less dubious and to study them in more detail to ascertain their nature and ontology. Wilson however does not provide much direction for this advancement of the understanding of the metaphysics of morality beyond broad claims about progress being possible once the biological nature of morality is realised. Further, to say anything non-trivial about moral metaphysics, it requires somewhat more explanation from Wilson as to how "removing ethics from the hands of philosophers" may be the best approach. Presumably doing so will not result in useful engagement with current and past philosophical literature and because of this miss many of the known philosophical difficulties or complexities involved.

For example, what kind of metaphysical insights might be revealed by looking at morality through a sociobiological or evolutionary lens? Perhaps it will show that evolution has produced various mental faculties in human minds (biological brains), and that the functioning of these faculties produces outputs that constitute the psychological and social phenomenon we call morality. If this were the insight that sociobiology will provide, then it would appear to be somewhat misleading to view this as a new kind of insight, as this is more or less the same kind of goal many moral philosophers have been aiming at for most of the twentieth century: providing an adequately naturalised account of the phenomena of moral philosophy.

[91] *Ibid*., pp 143-148.
[92] *Ibid*., pp. 196-199.

The program of Naturalism in this sense (although of course the term is used in a number of different ways in philosophy), can be divided roughly into two parts: ontological and methodological. The ontological component consists in asserting that reality is constituted of only of those kinds of things that are studied by the natural sciences, and that 'supernatural' entities do not exist (ghosts, spirits, or gods and the like). So, attempting to show that morality fits into a scientifically respectable ontology and has an explanation (evolution) that does not posit unacceptable entities or causes, fits perfectly within this already existing philosophical program of naturalism. The methodological component of naturalism is concerned with a commitment to the scientific method as our best tool for discovering what reality is like. Again, like the ontological component, philosophers that are sympathetic to naturalism are likely to share this goal: the phenomena in question will be best investigated (descriptively) through the appropriate scientific disciplines – the social sciences, psychology, neuroscience, sociobiology, anthropology and so on.

Thus, if the purpose of is simply to show that morality fits within a reality that contains nothing 'supernatural', then this is a program that many philosophers have long been committed to, and Wilson would need to provide more direction as to how sociobiologists would be better equipped to pursue that program than those already engaged in the endeavour. Indeed, any such task will require engagement with already existing literature on the metaphysics of morality and other areas of metaethics already devoted to dealing with how morality fits into the world. Simply recognizing the naturalness of the genealogy of our moral tendencies does not provide answers to the interesting questions about the metaphysics of morality.

Wilson's thought that progress may be made in philosophy by recognising the biological bases of morality is not without precedent. It is commonly thought that some progress in philosophy is made through the emergence of scientific disciplines that make use of newly developed approaches and methodologies to address questions that were previously the domain of philosophy. As the questions become systematically and conclusively answered, they move out of the purview of philosophy into

their own new fields of study. For example, 'natural philosophy' originally subsumed all kinds of systematic study of nature. During the scientific revolution, as new standards of evidence, experimentation, and methodology were developed and applied to particular problems, the study of natural philosophy as a discipline disappeared and was replaced by the familiar disciplines of biology, physics, astronomy, chemistry and so forth.[93] While such terms were already in use, they came to be more narrowly and rigidly defined to the refer to the application of scientific methods to these areas of inquiry.

Under this paradigm, once a methodology which conclusively answers questions within some domain of philosophical investigation is developed, the problem ceases to be a philosophical one. There are, however, limits to this process, and it appears highly unlikely that all of philosophy is amenable to such disciplinary speciation. David Chalmers has proposed a number of distinctive features that philosophical questions have that explain why some questions may remain as divisive questions that are part of philosophy rather than developing into new fields of science.[94]

The first feature that marks questions as philosophical is that they simply may not have objective answers. Where there is no objective truth, there cannot be a methodology that can consistently converges on that truth. A science where there is no underlying shared subject is an untenable endeavour. This explanation is only a partial one however, as not all areas of philosophy are susceptible to anti-realist theories.

Another feature of philosophical questions Chalmers suggests is what he terms 'verbal disputes', where proponents of opposing views use the same terms in different ways. While participants might be making sound arguments, they fail to engage properly with each other, and therefore gain little traction in persuading one another. Another possibility is that philosophy deals with domains that are remote from data that might answer the questions being asked. However, 'philosophical data' is not

[93] David Cahan, *Natural philosophy to the sciences: writing the history of nineteenth-century science*, p. 4.
[94] Chalmers, David, 'Why isn't there more progress in Philosophy?', pp. 20-21.

a commonly used term in philosophy and it is not clear what kinds of considerations would count as such. This worry about the remoteness of data could simply be another way of stating that empirical considerations do not appear to be relevant or sufficient by themselves to answer many of the questions which philosophy addresses. At best it is an unusual styling of philosophical methodology, and hints perhaps at the inappropriateness of speaking of 'data' in reference to the kinds of questions philosophy deals in.

Sociological explanations almost certainly play some role in ensuring philosophy's questions remain divisive. This may be because disagreement provides higher rewards in philosophy compared to the sciences, or that philosophical positions are simply more powerfully influenced by surroundings and the intellectual environment in which they develop than scientific positions. A striking example of the latter is G. A. Cohen's reflections on how, in a very real sense, he came to accept and defend the analytic/synthetic distinction due to attending Oxford, while recognising that, had he attended Harvard, he would likely have rejected the distinction, independent of the actual reasons for or against it.[95] Chalmers suggests that the other proposed features of the philosophical domain also give sociological factors more traction, thus making them more powerful in combination. Closely related to sociological explanations are psychological factors; there might be something unique about our minds and their self-reflexive relation to philosophical questions that makes us less able to converge on agreement in the way sciences might require.

All of the above considerations could be applicable to metaphysics of morality, and so provide reason to be wary of Wilson's claim. This is not to say that the direction of his thought is without merit. As I

[95] Cohen writes "...people of my generation who studied philosophy at Harvard rather than at Oxford for the most part reject the analytic/synthetic distinction. And I can't believe that this is an accident. That is, I can't believe that Harvard just happened to be a place where both its leading thinker rejected that distinction and its graduate students, for independent reasons—merely, for example, in the independent light of reason itself—also came to reject it. And vice versa, of course, for Oxford. I believe, rather, that in each case students were especially impressed by the reasons respectively for and against believing in the distinction, because in each case the reasons came with all the added persuasiveness of personal presentation, personal relationship, and so forth. So, in some sense of "because," and in some sense of "Oxford," I think I can say that I believe in the analytic/synthetic distinction because I studied at Oxford. And that is disturbing. For the fact that I studied at Oxford is no reason for thinking that the distinction is sound." In G. A. Cohen, 'Paradoxes of conviction', p. 18.

discuss in the following chapters, there are philosophers who think that evolution and biology have implications for metaethics, but to reach any conclusions such implications will likely require philosophical work. The direct application of scientific methods to metaethical questions will likely flounder due to the features highlighted by Chalmers. Only through looking at the actual metaphysical and metaethical debates involved, and through close engagement with precisely what the purported problems of metaphysical strangeness or super-naturalness are in ethics, is it likely that any progress will be made. And certainly, Wilson does not provide reason to think that this work would be best accomplished by having ethics "removed temporarily from the hands of the philosophers and biologicized."[96] I look at two such attempts to integrate metaethics with research from psychology and evolutionary biology in §3.2 and §3.3, but regarding Wilson's claims, very little simply 'falls out' of the recognition that morality or a tendency to moralize has a biological basis.

3.1.2 The problem of altruism

Wilson's second claim about biologicizing ethics is that evolutionary theory solves the 'problem of altruism'. The various processes discussed in the first chapter, including kin selection, reciprocal altruism, group selection, and mutualism, all show how behaviour that is individually costly to an organism could still be adaptive and therefore evolve. Further, sensible accounts have been given of extending such processes and adding others to them such as cultural evolution and cumulative cultural inheritance, to explain the emergence of morality in humans, and thus altruistic action in humans. Wilson's target with this claim about solving the 'problem of altruism' is the thought that altruistic behaviour would "work against individual survival – the altruist increases the chances of others' surviving, by helping them, while at the same time decreasing the chances of his own survival, by giving something up. Therefore, we would expect natural selection to eliminate any tendency towards altruism." However, the difficulty with Wilson's claim is that to a large degree, this "problem of

[96] E. O. Wilson, *Sociobiology: The new synthesis*, p. 562.

altruism" was only a problem of biology, not moral philosophy, and the difficulty of explaining altruistic behaviour prior to the discovery of such processes was a problem of biological theory.

There has in philosophy been much written on whether we are fundamentally egoists or altruists in the non-biological sense, that is, whether one is always motivated by one's own interests as opposed to acting in others' interests at a cost to one's own (and where interests is not defined as the interests of one's genes' interests). Practically every system of morality involves recommending actions that help others, often when such actions incur costs to the individual. However, there is one historical view of human nature that claims that we are incapable of following such imperatives, and that it is simply human nature to be selfish. This theory, typically called "psychological egoism" (or sometimes shortened simply to "egoism") holds that each individual looks out for only themselves, and thus they are unlikely to act altruistically: they will only help others when doing so is in their interests. Thus psychological egoism might be considered a "problem of altruism" in ethics. There are however a number of definitive responses to this view, and these responses were produced without the help of evolutionary theory.

In brief, a typical response is as follows. It is easy to reinterpret people's motives as being purely egoistic, for example by claiming that the reason we help people is so that we do not feel guilty, or because we think others will think more highly of us if they see us or hear about us doing so. While it is possible to reinterpret the motives that people have in this way, doing so does not really show that this is why people are truly motivated. One major reason is that such reinterpretation of motivations is very difficult to verify or falsify: any motive given for action will be translated into a self-interested one, meaning that no motive could be given for an action that would falsify the theory. Further, even if it were possible to establish that people were motivated because of these self-interested reasons, acting out of self-interest is not mutually exclusive with genuinely having the interests of others at heart – it is possible that pursuing one's own interests involves furthering the interests of others. Indeed many cases of everyday behaviour that involves helping others are just like this: for example

when people help their children or friends in any number of ways – they may be acting because it is what they want to do (what they consider 'in their interest'), but this fact does not show they do not genuinely wish to advance the interests of their children or friends that they are helping. Further, if some given action turned out not to advance (say) their children's interest, it is likely that they would not wish to act in this way at all.

So, a confusion that lends the appearance of plausibility to psychological egoism is the conflation of self-interested desires with selfish ones. Doing something out of self-interest does not necessarily preclude that action from being altruistic (in the general, everyday sense): self-interested actions differ in an important respect from selfish actions. Selfish actions actively disregard others' interests, whereas an action that is motivated by self-interest does not necessarily make any impact upon others (brushing one's teeth for example is done with one's own long-term interests in mind, but it need not be to the detriment of others!) Thus, once we focus on only actions that are selfish, it is much less plausible that this class of actions makes up the sole motivation for people's behaviour. So, there are clear reasons for rejecting psychological egoism, which do not depend upon biological explanations.

It is clear then that sociobiological insights are not the only way that this debate may be settled, as many philosophers have come to satisfactory and widely accepted conclusions that we do sometimes act in the interests of others at costs to ourselves. Therefore, given that Wilson's target appears in the first place to be one of biology (the problem of biological altruism), and that analogous problems in ethics have fairly widely agreed upon solutions that do not depend on sociobiological insights for their resolution, it would appear that Wilson's second proposal is somewhat unnecessary: ethics, at least in this debate, does not need assistance from biology.

3.1.3 Biological constraints on what we ought to do

Wilson's third proposal is the idea that paying close attention to our evolutionary origins may inform us of the limits of what is possible for humans: that doing so will show that our biology limits the range of forms that human nature may take. Morality recommends things that we should do, but there is little point to it if we are unable to put into practice what it recommends. Thus Wilson's idea is that evolution *constrains* the range of possible actions recommendable by morality, and sociobiology can improve our understanding of ethics by telling us what these constraints on what is possible are, due to our evolved human nature. There is a sense in which Wilson's second point about the problem of altruism (in the preceding discussion) could just be one example of this: if humans turned out to be fundamentally selfish, and incapable of altruism, then morality's dictates would be of little use. It could not be the case that we ought to be altruistic, if our evolved nature constrained our desires to only those that were selfish. So, the basic idea here is that "a sound morality must be based on a realistic conception of what is possible for human beings."[97]

Wilson suggests the following case as an example of a constraint on what is morally possible for humans due to our biology:

> Now there is reason to entertain the view that the culture of each society travels along one or the other of a set of evolutionary trajectories whose full array is constrained by the genetic rules of human nature. While broadly scattered from an anthropocentric point of view, this array still represents only a tiny subset of all the trajectories that would be possible in the absence of genetic constraints. As our knowledge of human nature grows, and we start to elect a system of values on a more objective basis, and our minds at last align with our hearts, the set of trajectories will narrow still more. We already know, to take two extreme and opposite examples, that the worlds of William Graham Sumner, the absolute Social Darwinist, and Mikhail Bakunin, the anarchist, are biologically impossible.[98]

[97] *Ibid*., p. 70.
[98] E. O. Wilson, *On human nature*, p. 208.

Thus, Wilson thinks that at least these two ideals of social reform are outside the gamut of what is possible for humans, due to our particular human nature. What should we make of this claim? Certainly, it seems as though sociobiology might be suggestive of the kinds of social arrangements that creatures such as ourselves would find rewarding. Social creatures with motivational systems that find rewarding friendship, cooperation, reciprocity, justice, desert, and all the benefits that these bring in terms of allowing people and societies to cooperate and flourish, are likely to find systems such as anarchy, where some of these goods are unavailable, or Social Darwinism[99], where injustice and apparently arbitrary imperatives about the above are made, to be difficult, and perhaps unrewarding. But there is nothing that sociobiology has shown that makes any of these a biological *impossibility*.

While in political philosophy the term 'anarchy' refers to a rather broad range of political views, its essence is that no 'coercive institutions' are justified and that coercive institutions should be replaced by social and economic organizations based on voluntary contractual agreement. This hardly sounds like a biological impossibility; indeed, surely similar arrangements of living have been from time to time part of our history. And Social Darwinism, while it is based on a flawed interpretation of biological evolution, and produces its moral imperatives in a *highly* dubious way[100], does not appear to involve anything that is particularly taxing on what we know about human nature. If it were put into practice as a political movement, it would involve coercion of the weak in society by the powerful, but this is something that is hardly novel to human nature. Thus it is questionable how much work can be done by Sociobiology in Wilson's claims about Anarchy and Social Darwinism. While biology may be somewhat suggestive of how well any given social reform will fit with human nature, it is hardly decisive, and often only in hindsight is the 'fit' able to be seen. Thus, it would seem at least in these

[99] Social Darwinism was the idea, popular in the late 19th and early 20th centuries, that humans are, or should be, subject to natural selection based on their social standing and circumstances and various other arbitrary features that were deemed "good" or "bad". According to this theory the poor and disadvantaged were considered weak, and the rich and able were considered strong, and therefore should be given preferential treatment. Social Darwinism was used to justify political conservatism, imperialism, and racism and to discourage intervention and reform.

[100] The errors that Social Darwinism makes are the same as those discussed in the §3.1.4.

broad, general cases of particular social ideas that it is hard to draw out any definite or practical implications. Richard Rorty sums up how we might respond to this claim about sociobiology: if we imagine that the Sociobiologists inform us that some proposed moral code or social reform is impossible for us to adopt;

> We have, they tell us, run up against hard-wired limits: our neural layout permits us to formulate and commend the proposed change, but makes it impossible for us to adopt it. Surely our reaction to such an intervention would be, "You might be right, but let's try adopting it and see what happens; maybe our brains are a bit more flexible than you think." It is hard to imagine our taking the biologists' word as final on such matters, for that would amount to giving them a veto over utopian moral initiatives.[101]

3.1.4 Naturalness and morality

The fourth claim of Wilson's is a claim about deriving the moral status of something from facts about its evolutionary origins - the fact that it evolved 'naturally' is used as a guide for claiming it is acceptable or normal in some sense. For example, Wilson claims that what he takes to be the 'biological purpose' of sex in humans – pair bonding – appears to "argue for a more liberal sexual morality".[102] Wilson thinks that the historical western view of sex (which he attributes to Christian and Jewish ideas, with the examples he uses coming from the Catholic Church) holds that the "primary role of sexual behaviour is the insemination of wives by husbands"[103], that any form of birth control outside of abstinence should be prohibited, and that all "'genital' acts outside the framework of marriage"[104] are abnormal, including masturbation which is an "intrinsically and seriously disordered act"[105]. Wilson thinks these ideas are misled, because the Church has a mistaken view of human nature:

> The Church takes its authority from natural-law theory, which is based on the idea that immutable mandates are placed by God in human nature. This theory is in error. The laws

[101] Richard Rorty, 'Born to be good'.
[102] E O Wilson, *On human nature*, pp. 141.
[103] *ibid*., p. 141.
[104] *ibid*., p. 141.
[105] *ibid*., p. 141.

> it addresses are biological, were written by natural selection, require little if any enforcement by religious or secular authorities, and have been erroneously interpreted by theologians writing in ignorance of biology. All that we can surmise of humankind's genetic history argues for a more liberal sexual morality, in which sexual practices are to be regarded first as bonding devices and second as means for procreation.[106]

Thus, Wilson appears to endorse the claim that because pair bonding is at least as much an adaptive primary function of sex in humans as reproduction, that it "argues for a more liberal sexual morality" – that it should be permissible to have sex for reasons other than reproduction.

However, it does not follow that something is good or right from a claim about something being natural or adapted. This form of argument, sometimes called an "appeal to nature" has a number of problems. Firstly, it is easy to show that if we allow this form of argument, there will be countless counterexamples: it is trivial to find things that are natural but are considered bad or things that are unnatural but considered good. For example, naturally occurring poisons and poisonous plants are often considered bad, but are natural, and vaccinations that produce immunity to diseases are unnatural, but are considered good. Evolved behaviour in all organisms includes the full range of what we consider morally good and bad behaviour, and thus it would be impossible to draw useful conclusions about good and bad just from the status of something as being an adaptation or not. Secondly, a more general error that 'appeals to nature' commit is that they jump from one descriptive fact, to a normative claim with no supporting (normative) premises in the argument.

At times Wilson appears to be aware that his argument is on shaky ground. For example, he provides a more careful discussion after coming to the conclusion that homosexuality could be a "naturally" occurring behaviour:

> The juxtaposition of biology and ethics in the case of homosexuality requires sensitivity and care. It would be inappropriate to consider homosexuals as a separate genetic caste, however beneficent their historic and contemporary roles might prove to be. It would be even more illogical, and unfortunate, to make past genetic adaptedness a necessary criterion for current acceptance. But it would be

[106] E O Wilson, *On human nature*, pp. 141-142.

> tragic to continue to discriminate against homosexuals on the basis of religious dogma supported by the unlikely assumption they are biologically unnatural.[107]

But of course, even this misses the point: the issue is not whether the practice in question *is* "biologically unnatural" or not – for whether it natural or not is not determinative or even indicative of whether something is generally considered right or wrong in philosophy (at least without significant further argument).

Thus, Wilson's claim that we should consider the "possibility that the time has come for ethics to be removed temporarily from the hands of the philosophers and biologicized" appears to be unfounded given the lack of substance of the four proposed routes of 'biologicization' that he presents. Biologists are unlikely to have special insight into the consequences for the metaphysics of morality based solely on more detailed understanding of the biology of sociality. Wilson's 'problem of altruism' turned out to be a problem of biology, not ethics, and the analogous problem in ethics – the idea of psychological egoism – is a problem to which there are already a number of apparently successful philosophical responses. While Wilson's idea that we must "adjust our ethical judgments to fit the realities" of human nature has merit in principle, sweeping claims about various political theories are too unspecific to be of much use. Claims about psychological or human impossibility will need to be very specific and will require philosophical work to establish, and the simple fact that some human behaviour or trait is evolved does little to show that it is impossible or even difficult to change. And finally, claims about the naturalness or unnaturalness of various behaviours or traits, do not tell us about whether they are morally good or bad: natural things can be bad, unnatural things can be good.

Thus Wilson's remarks appear to be overly optimistic about the ease of which biology and evolution science can be simply applied to revolutionise moral philosophy. This is not to say that it is impossible however or destined to be fruitless. Attempts such as Wilson's however, do point to the fact that it will require careful engagement with the philosophical debates if progress is to be made. To say

[107] E O Wilson, *On human nature*, p. 147.

something useful about ethics or metaethics, one must engage with the already existing ethical and metaethical debates. In the next section I look at one such attempt to do so by Richard Joyce, who argues that evolution has implications for the epistemic status of our moral beliefs.

3.2 Richard Joyce's evolutionary debunking argument

In recent years Richard Joyce has developed an evolutionary debunking argument.[108] Joyce's argument has been revised over the years, but it has remained a debunking argument in that that it has taken evolution to be showing certain pre-suppositions or commonly accepted assumptions about morality to be false. Based on considerations about the evolutionary origins of morality, at various points, Joyce has argued that different conclusions or consequences follow from his argument, including error theory, moral scepticism, and scepticism of the justification of moral beliefs. It is instructive to see how his argument has developed, as it is a good example of an argument that has set out to see what follows from the evolutionary origins of morality, while not neglecting to engage with the meta-ethical literature and taking into account considerations the literature raises. The end result is a mature position that is neither philosophically naïve nor under-developed and is surprisingly robust given its potentially controversial outcomes.

His final position is that at the least, our moral beliefs' *justification* is undermined by our understanding of the evolutionary origins of morality (it may be that we can re-instate such justification, but not without providing an argument independent from genealogical considerations). His conclusion is an epistemological one; the argument aims to show the justification for moral beliefs is lacking or that it never existed all along. It is worth noting that this conclusion is weaker than the conclusions of some other arguments that have been termed "evolutionary debunking arguments"[109]

[108] Richard Joyce, *The myth of morality, The evolution of morality.*
[109] For examples, see Guy Kahane, 'Evolutionary debunking arguments'.

and the resulting meta-ethical position does not by itself, constitute an error theory, as its conclusion is still consistent with there being true moral beliefs.

The starting point of Joyce's argument is the observation that the best accounts we have of the evolution of our moral psychology seem to show that its evolutionary function is to facilitate social cohesion and group living. By evolutionary function it is meant the reason that our moral psychology was selected for; why it made those particular individuals or groups[110] who possessed it more reproductively and therefore evolutionarily, successful. If this is the function of our Moral Psychology, then the reason it evolved does not necessarily rely on the moral beliefs it produces being true. The metric for success in the evolutionary environment was that the beliefs were *useful* rather than true, which raises the question of where moral truth fits in this genealogical story, and if it does so at all.

3.2.1 Belief Pills

Joyce often introduces his argument via the following thought experiment. He asks us to imagine there are such things as 'belief pills' which cause us to start believing something when we take them but to have no memory of this belief formation process. Say that you believe that Napoleon lost the battle of Waterloo. Sometime later you learn that the reason you have this belief is that someone tricked you into ingesting a belief pill that made you believe that he lost at Waterloo. To aid the arguments in this chapter I will use basic flow charts of the following kind[111] - the simple belief pill case is as follows:

[110] Depending on which account of the levels of selection at work in the evolution of morality one subscribes to.
[111] The nature of the entities and relations in the diagrams will not be specified – they are not necessarily causal, identity, entailment, implication or any other kind of formal relation. Their purpose is simply to allow the reader to more easily follow the flow and order of the arguments.

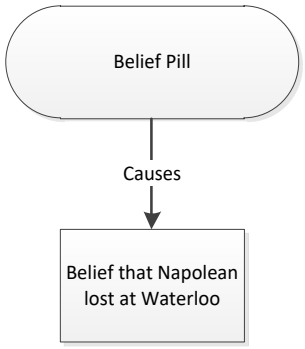


Figure 4: Belief pills

Given the new knowledge about the origins of your belief about Napoleon losing at Waterloo, you have a choice of what attitude to take towards this belief. You could carry on believing that Napoleon lost at Waterloo, but doing so seems remiss: you know that the truth of the belief has no connection to why you believe it, so why should you continue to think that it is true? You could adopt the belief that it is false that Napoleon lost at Waterloo, but this too seems to be an inappropriate response; your new knowledge about the belief pill provides no more reason for you to believe it to be false than it does for you to believe it to be true. In this position of uncertainty, it would seem best, at least initially, to suspend judgement about the truth or falsity of Napoleon losing at Waterloo and hold that your previous belief is no longer *epistemically justified*.

A similar line of reasoning can be applied to the case of the evolutionary genealogy of morality, where the process of the evolution of our disposition to make moral judgments is like the belief inducing pills and our moral beliefs the belief about Napoleon losing at Waterloo. In applying this analogy, there is one modification that is immediately necessary for the analogy to go through. The hypothetical belief pills, as their name suggests, produce beliefs. Evolution however, does not simply produce innate beliefs in humans about what is right or wrong (for example, I can have a belief that I morally ought to pay my taxes, but it is unlikely that the forces of natural selection at play could have 'known' about government revenue and redistribution of income). The connection between the evolutionary genealogy to the particular moral beliefs has an intermediary step. Instead of directly producing

beliefs, the result of the evolutionary genealogy is a psychology that judges the interactions and events of social life in terms of moral concepts and forms beliefs that these moral concepts figure in.[112]

The belief pill analogy can be modified to account for this intermediary step, such that instead of producing particular beliefs (that Napoleon lost at Waterloo), the pills cause the taker to form beliefs involving a particular concept; that is beliefs involving the concepts of battle, Napoleon, Waterloo and so on in some sensible combination. Without taking the pill you would not form beliefs about Napoleon at all. The exact content of the belief is irrelevant to the fact that were you to discover that someone had tricked you into taking a Napoleon-type belief pill, your justification for holding the belief about Napoleon would be undermined.

Thus modified, with evolution as the substitute for belief pills, causing humans to form beliefs involving particular concepts; beliefs involving rightness and wrongness, justice, fairness, and so on. The best accounts of the evolution of morality we have indicate our moral psychology developed because it was useful rather than any other reason (for example because it successfully identified true beliefs). Without our particular evolutionary genealogy, we would not form beliefs involving these moral concepts at all. Joyce thinks that once we have uncovered these facts about evolution and morality, we should seriously question whether our moral beliefs are appropriately justified and should be sceptical of believing those moral beliefs are true. Regardless of the truth-values that we think our moral beliefs have, our justification for holding these beliefs should be undermined upon discovering the evolutionary origins of the concepts that figure in and constitute them.

[112] There are a number of different ways this could be developed. Joyce discusses these in more detail in 'Evolution and ethics'. One of these is that the human brain comes hardwired with moral concepts but the developmental environment determines what things the concepts are applied too. Thus, in one environment individuals may develop to judge that slavery is deeply morally wrong, whereas in another it may be viewed as entirely acceptable. Another view is that the concepts the moral faculty deals in come with biases towards certain kinds of content. Cross-cultural surveys show that the moral concepts are relatively universally deployed in domains that deal with actions producing harm, regulations concerning fairness and exchanges, values pertaining to social hierarchy, and so on. Thus, moral systems involving these domains will be more easily learnt and perhaps there are some prewired abstract principles and innate parameters which when combined with a particular developmental environment result in a fully functional moral faculty. Examples of this kind of conception are elaborated in Marc Hauser, *Moral minds,* and John Mikhail, *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*.

The diagram below shows the structure of this proposed model of the influence of evolution on moral beliefs:

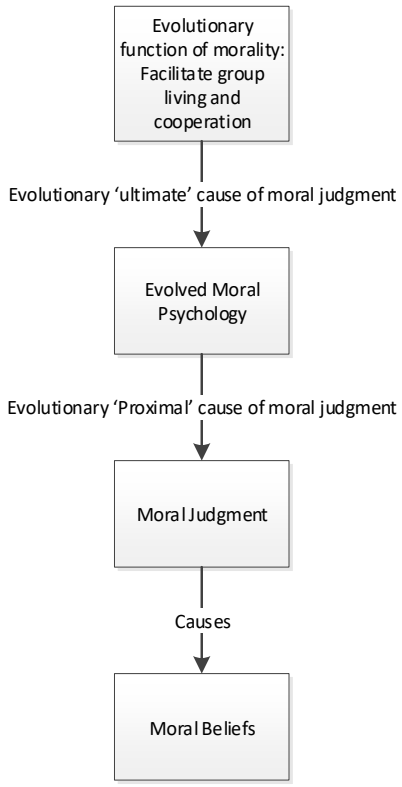


Figure 5: Evolutionary genealogy of moral belief

While it is possible that the moral beliefs are true, the above picture gives moral truth no role in the causation of moral beliefs either at the proximal or ultimate levels.[113] Thus this picture gives us no reason to believe that our moral beliefs are likely (or unlikely) to be true and instead it shows the reason we hold these beliefs at all is one that is unrelated to their truth, and thus we should re-

[113] This refers to Nikolaas Tinbergen's levels of explanation of animal behaviour which draws the distinction between Proximate explanations which explain how organisms work by describing their structures, mechanisms, and ontology versus Ultimate explanations which explains why organisms are the way they are by describing how selection shaped their current form. Originally described in Nikolaas Tinbergen, 'On aims and methods of Ethology'.

consider their status as epistemically justified, unless we can discover some other source of justification.

3.2.2 Truth tracking

The obvious response for those who wish to save the epistemically justified status of our moral beliefs is to claim that moral truth did in fact have a role in the evolutionary development of our moral psychology in some manner. In the context of evolutionary debunking arguments, the term *truth tracking* is often used to refer to the doxastic relation between our evolved moral psychology, and the moral facts which the moral beliefs are purportedly about. Our moral psychology is said to *track the truth* when the beliefs it produces are connected in an appropriate manner with the truth of moral facts. By appropriate manner, without going in to the details of any particular epistemic theory, it is simply meant that to find out that this relation does not hold is to discover that the belief is not appropriately justified or backed up by the way things are.

Thus, a truth tracking account of the evolution of morality would require that the moral psychology that was selected for had some way of recognising or "latching on" to the moral truth that the content of the moral belief was about. Evolution would be selecting creatures that could accurately pick up on what the moral truth in a given situation was. The following picture shows how a simple version of a truth tracking account of the evolution and functioning of morality might look:

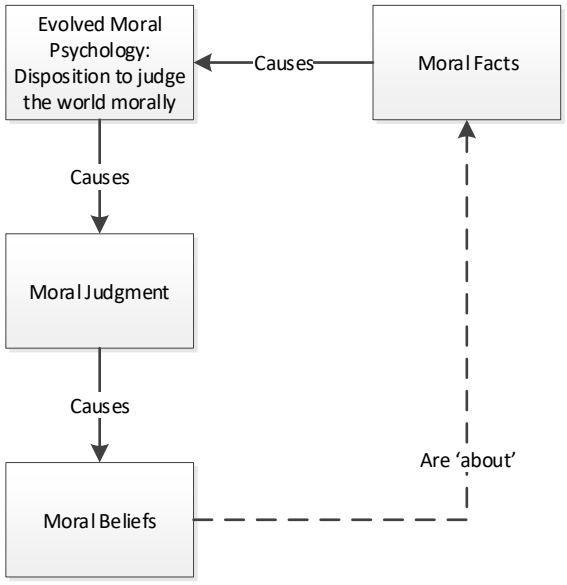


Figure 6: A truth tracking account of morality

However, is this picture of the evolution of morality a defensible one? Could evolution have evolved to latch-on to and track independent moral facts? Joyce notes that it may be objected that the Napoleon thought experiment is rigged from the start to undermine the beliefs in question: if beliefs are formed in a manner that does not relate to the truth or falsity of the proposition, then of course knowledge of this fact undermines their justification. Perhaps morality is not like this, and evolution has a tendency to produce a disposition to form true beliefs about morality.

3.2.2.1 A truth-dependant evolutionary genealogy

To make sense of this, Joyce draws a distinction between two kinds of evolutionary explanation for why people have the beliefs that they do. The first kind are explanations that depend upon the truth of the beliefs they're explaining for them be successful explanations of why such beliefs bestowed evolutionary advantages. The example he gives of this is the human faculty for doing simple arithmetic. He claims that there is (or could be) a straightforward evolutionary explanation for why humans would have an inbuilt faculty that allows them to perform basic calculations. Does, in the case

of arithmetic, an evolutionary genealogy risk undermining our beliefs about simple propositions of arithmetic? Joyce does not think so:

> …let's interpret this as implying that our belief that 1 + 1 = 2 is innate. This, it seems pretty safe to declare, is an eternal and necessary truth, and thus by "hard-wiring" such a belief into our brains natural selection takes no risks—it is not as if the environment could suddenly change such that 1 + 1 would equal 3. So does the fact that we have such a genealogical explanation of our simple mathematical beliefs serve to demonstrate that we are unjustified in holding these beliefs? Surely not, for we have no grasp of how this belief might have been selected for, how it might have enhanced reproductive fitness, independent of its truth. False mathematical beliefs just aren't going to be very useful. Suppose you are being chased by three lions, you observe two quit the chase, and you conclude that it is now safe to slow down. The truth of "1 + 1 = 2" is a background assumption to any reasonable hypothesis of how this belief might have come to be innate.[114]

So, in this case, according to Joyce, the evolutionary genealogy is not an undermining one, as the availability of fitness benefits depend upon the beliefs in question being true. Justin Clarke-Doane however, has argued that Joyce's interpretation of the example is incorrect[115], despite the view being widely accepted by many, including Roger Crisp, Allan Gibbard, Stephen Pinker, Walter Sinnott-Armstrong, and Ernest Sosa.[116] Clarke-Doane argues that this example is mistaken because it construes what are first order logical truths as mathematical truths. Clarke-Doane asks us to imagine two hypothetical ancestors, *P* and *Q,* both of who observe a pair of lions entering some bushes nearby to hide. The difference between the two is that *P* believes there are two lions entering the bushes to hide and that 1+1=2, whereas *Q* while also believing there were two lions entering the bushes to hide believes that 1+1=0. He argues that ancestors like *P* would have had an evolutionary advantage over those like *Q*, not because of their beliefs about arithmetic, but because the logical truths corresponding to those arithmetical statements obtained. That is:

---

[114] Richard Joyce, *The evolution of morality*., p. 182.
[115] *Ibid.*
[116] Clarke-Doane references Roger Crisp, *Reasons and the good,* Allan Gibbard, *Thinking how to live,* Steven Pinker, *The Blank Slate,* Walter Sinnott-Armstrong, *Moral Skepticisms,* and Ernest Sosa, 'Reliability and the A Priori', see Justin Clarke-Doane, 'Morality and mathematics: the evolutionary challenge', p. 327.

> If our ancestors who believed that 1+1=2 had an advantage over our ancestors who believed that 1+1=0, the reason that they did is that corresponding (first-order) logical truths obtained. In particular, ancestor *P*, who believed that 1+1=2, had an advantage over ancestor *Q*, who believed that 1+1=0, in the above scenario [where *P* believes there were two instances of lions entering the bushes to hide, and that 1+1=2 vs *Q* who also saw the lions entering the bushes but believes that 1+1=0], intuitively because if there is exactly one lion behind bush *A*, and there is exactly one lion behind bush *B*, and no lion behind bush *A* is a lion behind bush *B*, then there are exactly two lions behind bush *A* or *B*. In other words, ancestor *P* did not have an advantage over ancestor *Q* because its belief that 1+1=2 was true. Ancestor *P* had an advantage over ancestor *Q* because its belief appropriately aligned with (first-order) logical truths about its surroundings.[117]

Clark-Doane's contention, is that beliefs such as 'there are exactly two lions chasing you, and you observe two lions quit the chase, so you conclude that it is now safe to slow down' are not constructions that refer to arithmetic and hence mathematical truths. Instead they are first order logical beliefs.[118] He also points out that whether a statement counts as a first-order logical truth as opposed to a mathematical one is not simply a terminological issue, and while many claim that mathematics reduces to logic, the claim is not that it reduces to first order logical claims, but to second-order logic or set theory.[119]

Clarke-Doane is likely correct about this point, but for Joyce's argument, the fact that the chased-by-lions vignette is about first order logical beliefs instead of mathematical beliefs is not detrimental. While the point may be of interest to philosophers of mathematics, acceptance of Joyce's claim that some beliefs may have that may have an evolutionary origin that does depend upon the truth of those beliefs all that is required for his argument. Indeed Clarke-Doane is willing to accept that such beliefs exist as long as we do not talk of them as mathematical beliefs.[120] Clarke-Doane also finds a number

[117] *Ibid*., pp. 330-331.
[118] Roughly formalised the belief would be that "there is an *x* and a *y* such that *x* is a lion chasing you that has quit the chase and *y* is a lion chasing you that his quit the chase and, x ≠ y, and for all *z*, if *z* is a lion chasing you, then *z* = *x* or *z* = *y*".
[119] *Ibid*., p. 330, see footnote 41.
[120] *Ibid*., pp. 331-333. Clarke-Doane accepts many mathematical hypotheses correspond to nonmathematical truths about the world that could figure in evolutionary genealogies that depend upon the truth of the beliefs. He writes "the basic idea that for any mathematical hypothesis that we were selected to believe, *H*, there is a

of issues elsewhere in Joyce's argument, and evolutionary debunking arguments in general, which are discussed along with the arguments of William Fitzpatrick and Benjamin Fraser in further detail in §3.5.

3.2.2.2 Truth-independent evolutionary genealogy

The second kind of evolutionary explanation is one that explains why the beliefs we have a disposition to form, were fitness enhancing, regardless of whether these beliefs were true. Joyce thinks that the moral case differs from the arithmetic (or, accepting Clarke-Doane's contention, the first order logical) case in just this way: we can make sense of our ancestors' disposition to form beliefs about rightness and wrongness independently of the existence of anything that these terms refer too. The practical success that having such beliefs could have bestowed could be sufficient to ensure they were fitness enhancing regardless of whether they were true or false. Moral systems, in the very first stages of the evolution of morality could have been no more than primitive rules or emotions or conventions aimed at ensuring social cohesion and collective action. And at further stages, there is no obvious reason why this criterion of success (social cohesion) would need to be any different. Thus, beliefs that accurately represented facts were not necessary for the fitness benefits of greater social cohesion and cooperation to be available. Although truth is often a good route to practical success (for example when counting the number of lions chasing you), it is possible that having false beliefs and acting on them could also have been adaptive in certain contexts. As Joyce notes, "Whether we assume that the concepts of *right* and *wrong* succeed in denoting properties in the world, or whether we think that they suffer from a referential failure that puts them on a par with the concepts *witch* and *ghost*, the plausibility of the hypothesis concerning how moral judgment evolved remains unaffected."[121]

nonmathematical truth corresponding to *H* that captures the intuitive reason that belief in *H* was advantageous is plausible. By nonmathematical truth I mean a truth that does not imply a substantive mathematical sentence…that is, roughly, a truth that does not imply the existence of a relevantly mind-and-language independent realm of mathematical objects." p. 332.

[121] Richard Joyce, *The evolution of morality*, p. 183.

3.2.3 Moral naturalism

The first kind of evolutionary explanation described above shows that moral truth has a potential role it could play, but it does not explain why it was evolutionarily advantageous to track the moral truth or any of the details of how it might work. Why would the moral faculty have this seemingly ad hoc (from a scientific viewpoint) 'identify and track moral truths' function? Without a reason for introducing moral facts into the picture (other than to reply to the skeptic) there is much left unexplained about such a truth tracking account of the evolution of morality.

The most common way of resolving this problem is to endorse some variety of moral naturalism: claiming that moral truth was part of the evolutionary explanations all along. The things that our evolved moral psychology tracked were things that enhanced social cohesion and group living, and these just were, in some sense, the moral facts.

Moral naturalism is this view that moral facts are identical to, constituted by, or supervene on some natural facts. So, moral facts are simply facts about humans, their environment, and their "patterns of reaction to it"[122] and therefore are the kinds of things that can play a causal role in explanations. This view, if it can be defended, provides a potential truth tracking account of the evolution of moral judgment. The diagram showing how this position fits with the evolutionary account is as follows:

[122] Simon Blackburn, *Spreading the word: Groundings in the philosophy of language*, p. 182.

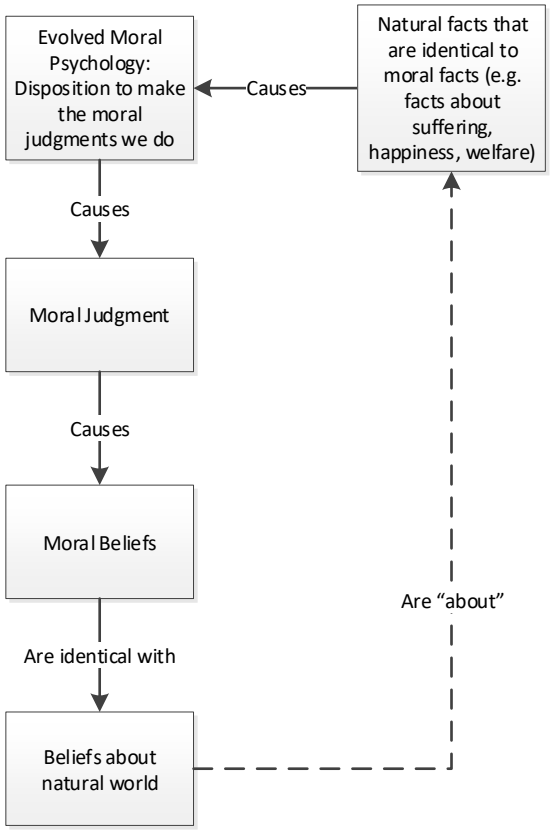


Figure 7: Evolutionary genealogy of moral naturalism

3.2.4 Harman's challenge

Joyce takes the evolution of morality as an explanation of why we have a disposition to form moral beliefs, and this explanation does not depend upon moral beliefs being true. Further, he claims that without this evolutionary genealogy and our particular social history, we would not have such moral concepts as obligation, virtue, wrongness, desert, fairness and so on at all. Therefore, if there is no explanation for why we would have these concepts independent of the evolutionary explanation, then it seems that we have reason to doubt that moral beliefs would be independently true or that the concepts that figure in them ever refer to anything. Recognition of this means that unless the moral-

realist can make good on their picture of moral naturalism, a skeptical conclusion looms. Joyce thinks this makes what he call's 'Harman's Challenge' particularly pressing.[123] Gilbert Harman's challenge is that if we can explain why we make a moral judgment without reference to moral facts, and we cannot show how in some sense moral facts figure in, reduce to, or supervene on the original explanation, then we should cease to think that the judgment has anything to do with these facts.

Harman uses the following example to illustrate this challenge:

> You round a corner and see a group of young hoodlums pour gasoline on a cat and ignite it, you do not need to *conclude* that what they are doing is wrong; you do not need to figure anything out; you can *see* that it is wrong.[124]

Suppose that we attempt to explain this event of judging the cat-burning in terms of the natural sciences. We can give an account in the terms of physics and chemistry say, in which the terms "cat", "burning", and "wrong" do not figure at all. But does the fact that these terms do not appear in the explanation mean they have no part to play in explaining this case of cat burning? For the terms "burning" and "cat" it seems the answer is no: burning cats can be *reduced* to descriptions of physics and chemistry, which shows that the burning cat was in some sense 'present' in the causal explanation of the situation. Does the same hold for the term "wrong"? Can we provide a reduction of the term "wrong" to naturalistic facts? Harman's challenge is that if we can explain why we come to judge this to be wrong, why it *seems* that this episode of cat burning is wrong, without having any idea of how the "facts of wrongness" reduces to or supervenes upon this explanation, then the actual existence of wrongness is not needed to explain anything about the situation, and we need not think it plays any part in this episode of cat-burning. Thus, Joyce thinks he has an explanation for why people would hold *any* moral beliefs or use *any* moral concepts, and argues that unless there is an account of moral facts which shows how they reduce to or figure in moral explanations, then we should conclude that we are in error holding moral beliefs: there are no moral facts. If morality is to be vindicated, it must

[123] *ibid.*, p.184, Gilbert Harman introduces his 'challenge' in '*The nature of morality*', pp. 3-10.
[124] *ibid.*, p. 4.

be because some kind of moral naturalism (the idea that moral facts reduce to or supervene upon natural facts) is true. Otherwise moral facts should be excised from the picture with a "swift slash from Ockham's Razor, since we have a complete explanation of moral judgment with no need to posit any extra ontology in the form of moral facts."[125]

Joyce argues that unless Harman's challenge can be met, our commitment to morality should be undermined, and we should consider that moral discourse constitutes some kind of mistake: we should be 'error theorists' about morality. An error theory is a theory about a domain of discourse that one believes to have been shown to make false claims.[126] Typically in meta-ethics, the term 'error theory' is reserved for positions which involve active disbelief – situations where we are sure that the subject matter is false. While Joyce at this earlier point did describe his position as a kind of error theory, his view is not quite as strong as typical positions that use the term and thus, he no longer uses this designation. He argued that the evolutionary explanation of our moral beliefs forces us to consider Harman's challenge, and if we cannot meet it, then he thinks a position of agnosticism towards morality is justified. Ultimately, he concludes that the prospects for a satisfactory account of such a moral naturalism that would meet Harman's challenge are not good, and thus endorses the claim that moral facts do not refer to anything: moral discourse is a kind of error. In the next sections I examine Joyce's grounds for concluding that moral naturalism is untenable.

3.2.5 Arguments against moral naturalism

The strategy Joyce adopts to argue against moral naturalism is a kind of *reductio ad absurdum* whereby if it can be shown that commitments entailed by moral naturalism are not possible, then

[125] *ibid*., p.188.

[126] What is meant by 'error theory' is often conveyed by way of example – typical illustrations include discourse about supernatural entities such as ghosts or witches, or theories about scientific entities or facts that have been proven to be misled, such as the fact that phlogiston resides in combustible materials or that the earth is flat. We hold an error theory about these cases because we think that they are predicated on errors: there are no such things as witches, ghosts, or phlogiston, and the world has been shown to be spherical, not flat.

some or all of those commitments must be rejected. Joyce identifies a central feature of moral thought and discourse – a "non-negotiable" commitment that cannot be possibly naturalized: "it is very hard to see how naturalistic facts could possibly provide *the inescapable authority* we apparently expect and require of moral facts."[127] He calls this particular feature of morality 'practical clout' – the combination of moral inescapability and moral authority. He thinks any satisfactory analysis of the concept of morality and how it is used will show that practical clout is a necessary feature. Joyce's argument expands upon those of John Mackie in Mackie's *Ethics: Inventing Right and Wrong.* Joyce agrees with Mackie, that in making a moral judgment someone is saying something that "is not purely descriptive, certainly not inert, but something that involves a call for action or for the refraining from action, and one that is absolute, not contingent upon any desire or preference or policy or choice, his own or anyone else's."[128] What Joyce terms "practical clout" appears to be similar to what Mackie calls "authoritative prescriptivity", "objective prescriptivity" or even simply the "to-be-pursuedness" or "to-be-doneness" that is somehow built into moral claims.[129] Practical clout's inescapability amounts to the fact that one cannot avoid a moral prescription simply by having or not having particular desires. For example, one cannot "evade the proscription 'Don't kill innocent people' by citing some special desires that makes it ok for one to do so ('But I *really* enjoy it!') or shrug off moral concerns by claiming a lack of interest in such values."[130] Thus according to Joyce, moral considerations are supposed to bind people irrespective of what their desires or interests are. It is important to note that whether these considerations are taken any notice of by the moral agent is a separate issue. Joyce is not arguing for a kind of 'motivation internalism' whereby a moral judgment is not counted as sincere unless it is recognized as providing motivation for acting in accord with the judgment. As he writes, this strange feature of moral judgments (practical clout) "resides not in

[127] *ibid*., p. 191.
[128] John Mackie, *Ethics: Inventing right and wrong*, p. 33.
[129] For example, Joyce tries to show that practical clout is important to people, by quoting Mackie talking about "objective prescriptivity" and how *it* is important to people. Joyce agrees that "this assumption [of objective prescriptivity] has been incorporated in the basic, conventional meanings of moral terms" – where this quote is from Mackie talking about the status of objective prescriptivity (*ibid*, p. 35).
[130] Richard Joyce, *The evolution of morality*, p. 192.

intrinsic motivation-engagement but rather *intrinsic action-guidingness.*"[131] However despite not adopting such an internalist view, Joyce does think moral considerations still have a particularly *strong* form of authority, one that cannot simply be ignored. Joyce thinks that if some considerations are to be counted as moral, they must provide reasons for action that cannot be ignored in one's deliberations about what to do – although the ultimate result might not be deciding on the outcome that the moral considerations favour.

For Joyce the authority of morality can be contrasted usefully with that of etiquette: while etiquette provides reasons that are inescapable (people who do not care about etiquette can still be said to transgress against its demands), the reasons it provides do not have "Genuine binding force over a person."[132] For example, a person who "speak[s] with his mouth full in order to stop a friend from eating a wasp" is transgressing a norm of etiquette, but the reasons etiquette gives in this case are not authoritative – the fact that speaking in such a situation with one's mouth full might be considered rude is not a consideration that holds any deliberative weight in such circumstances. As Joyce writes, when compared to normativity of the kind etiquette provides "it is often thought that morality requires something stronger and more authoritative."[133]

Joyce's argument about practical clout has strong similarities with John Mackie's "Argument from queerness."[134] Mackie divides his well-known argument from queerness into two parts: a metaphysical part – that "if there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe"[135]; and an epistemological part – that "if we were aware of them, it would have to be by some special faculty of moral perception or intuition, utterly different from our ordinary ways of knowing anything else"[136]. Joyce's argument against moral naturalism focuses only on the metaphysical considerations raised by

[131] *ibid.*, p.173.
[132] *ibid.*, p. 192.
[133] *Ibid.*, p. 193.
[134] J. L. Mackie, *Ethics: Inventing right and wrong,* pp. 38-42.
[135] *ibid*., p. 38.
[136] *ibid*., p. 38.

Mackie, which seems to be a reasonable simplification given that the epistemological argument is dependent on the metaphysical one in the sense that one only needs to posit an 'utterly different' way of knowing due to the fact that the "entities or qualities or relations" are themselves "of a very strange sort, utterly different from anything else in the universe."[137] Mackie thinks that if moral properties existed they would have to be of the form of categorical imperatives (that is there a sense in which they apply to all inescapably)[138] and be "objectively valid"[139] and that such a combination is "metaphysically queer".[140] Similarly, Joyce's term practical clout appears to identify the same features: inescapability (for Mackie that moral judgments have the 'form' of categorical imperatives) and moral authority (that moral judgments have "objective validity" and are "intrinsically action-guiding"). Mackie thinks this objective validity means that the truth of moral facts could not depend in any way on people's desires or ends: they would need to be true regardless of whatever ends, interests, or desires people had and they would need to be intrinsically "action-guiding" for all. Mackie suggests that Plato's Form of the good would be an example of what such metaphysically queer entities might be like: "The Form of the Good is such that knowledge of it provides the knower with both a direction and an overriding motive; something's being good both tells the person who knows this to pursue it and makes him pursue it."[141] Such "entities, qualities, or relations" simply could not exist as they are too metaphysically peculiar: they would have to guide action by some unspecified and apparently mystical mechanism. Both Joyce and Mackie's arguments hold that morality has features that simply cannot exist as they are too metaphysically "queer", and do not fit within a naturalistic picture of reality.

So, Joyce's complaint about moral naturalism is as follows. Moral facts, if they are to exist, need to provide us with reasons completely irrespective of our individual ends, and these reasons must be

[137] *ibid*., p. 38.
[138] J. L. Mackie, *Ethics: Inventing right and wrong,* pp. 27-30.
[139] *ibid*., pp. 30-35.
[140] *Ibid.*
[141] *ibid*., p. 40.

authoritative – they cannot be ignored in practical deliberations about what to do – the reasons that awareness of moral facts give us must be *strong* reasons. It appears that these desiderata – practical clout or in Mackie's case objective validity – are utterly different from anything that we know how to explain naturalistically, and they must require in some sense for the "universe to take sides". If the moral facts are true independently of people's desires, concerns, or ends, then something else somewhere in the universe must demand that something be done or not done, pursued or not pursued. Mackie and Joyce's arguments do seem to be persuasive on this light: there does not seem to be any way we can imagine such a "queer property" as practical clout or any kind of objective prescriptivity reducing to or supervening on the natural world we know.

3.2.6 Moral Naturalism and practical clout

There are two ways one might respond to Joyce's argument against moral naturalism that reject his view of practical clout – the inescapable practical authority that Joyce argues is a necessarily a feature of moral judgments. The first kind of reply is simply to claim that practical clout can be accommodated by a suitable moral naturalism, and to demonstrate how this might be so. The second kind of response is simply to deny that practical clout is a feature of morality. I shall begin with the first kind of response: with showing that practical clout can be suitably naturalized. Joyce assumes that attempting to locate practical clout will mean identifying a particular kind of *reason* that people have when there is a moral requirement to do or refrain from doing something. This is the route that many supporters of moral naturalism take to be most promising and it is hard to see how else practical clout can be naturalized if not in the form of reasons, as talk of reasons is our best conception of how the notion of "to-be-doneness" or "to-be-pursuedness" is cashed out in.

3.2.6.1 Moral naturalism with practical clout

There is one kind of reason that is typically considered but rejected: those that are sometimes called "institutional reasons" following Mackie's discussion.[142] Suppose that I am playing chess, and I wish to move one of my Rooks diagonally. It appears acceptable to say, that even though I desire to do so, I have a reason *not* to move my Rook diagonally. The fact that I have a reason follows from the fact that I agreed to play chess (and thus tacitly perhaps, assent to playing by the rules). So, in this case I have a reason to act (or refrain from acting) that does not depend on my desires. Because of this, Mackie does not deny that there are "desire-transcendent" reasons. He does however reject that such reasons can be "objectively valid": such reasons are only legitimate because of the presence of an institution that one must endorse for the reason to apply. While this kind of reason is independent of my immediate wants, it is not entirely independent of all my desires or purposes: I only have a reason not to move my Rook diagonally if I accept the rules of chess. So, the reason only applies hypothetically. If I no longer wish to partake in the chess game, then I would no longer have a reason to play by the rules. These so called "institutional reasons" apply wherever there is any kind of established institution or social practice. Other examples that Mackie suggests are etiquette and the institution of promising.[143] Because these kinds of reasons are escapable in some sense, they are unable to provide the kind of practical clout necessary to be the kind of reasons morality gives us. If institutional reasons are ultimately dependent upon the ends and desires of the person, then moral reasons cannot be a kind of institutional reason. Because the required naturalization must account for reasons that are *inescapable*, they cannot be sensitive to peoples' wishes to partake or not partake in a particular institution.

[142] *ibid*., pp. 64-73.

[143] Although it is not clear that one can simply "opt out" of the institution of promising, Mackie thinks that there is a sense in which promising *is* optional or hypothetical: "I can surely refrain from endorsing the promising institution; I can decline to speak within it. No doubt this would be eccentric, unconventional, it might well make people distrust or dislike me, but it is not logically ruled out." *Ibid.,* p. 69.

Where else then might we find practical clout? One way is to locate the property of goodness as being identical to or supervening upon some other property, such as *the tendency to produce happiness,* or *the tendency to promote welfare*.[144] Certainly, it seems plausible that considerations of happiness or welfare might provide us with reasons for action (hence the intuitive appeal of utilitarian theories – it often just seems that suffering should be avoided, happiness promoted where possible, and so on). But will awareness of these properties provide reasons that are inescapable and authoritative? It seems hard to deny that welfare and happiness and the like are things that generally matter to us, but does the fact that an action will produce happiness or welfare (or whichever property we choose) *always* produce reasons in us that have deliberative weight? Or, is the fact of the matter simply just that while we generally do think that such things give us reasons, there are cases where we can be aware that such a property as *tendency to promote welfare* obtains, but that this fails to make a difference to our reasoning? It seems that despite the attractiveness this proposal has, it is does not go very far towards showing how such properties themselves are inescapable and authoritative. There is no obvious way to show that there is a necessary connection between the property of *tending to promote welfare* (or whichever property we choose) and our reasons. This strategy therefore simply pushes the problem back a step, as we still need to answer the question: why think that the property of promoting happiness or welfare will have practical clout? As it stands, there is no argument other than the proposal's initial plausibility to show that such reasons are inescapable (that one can be aware of such a property, but that this has no impact on one's deliberation) and thus this is not a good solution to the problem of naturalizing practical clout.

Another way of approaching the naturalization of morality that is often thought to be more promising is to identify moral requirements as requirements of rationality: we are morally required to do just what we have sufficient or real reason to do. What we have real reason or sufficient reason to do is just what we would do if we 'reasoned correctly' in some sense. Here the strategy is to attempt to

[144] It is worth noting such a strategy is not limited to properties historically identified by utilitarians – those concerning happiness, welfare, pleasure and so on.

provide a "substantive and naturalizable account of 'correct practical reasoning' (or 'practical rationality') according to which any person, irrespective of her starting desires would through such reasoning converge on certain practical conclusions that are broadly in line with what we would expect of moral requirements."[145] The aim therefore is to find an account of reasoning whereby if someone has a genuine moral reason for acting in some way, call it ϕ-ing, then it is not possible for them to sensibly say something like "I acknowledge that were I to reason correctly I would want to ϕ, but what is it to me?"[146] According to Joyce, the "dominant attempt in modern philosophy" to provide such an account of reasoning is what is often called the 'self-conception strategy'. Joyce quotes David Copp who describes this strategy as follows:

> On the self-conception strategy, there is a way of conceiving of oneself such that a rational person who is thinking clearly *must* conceive of herself this way, but if she does not treat moral reasons as authoritative, she cannot *coherently* conceive of herself in this way. It might be said, for example, that a person who does not treat moral reasons as authoritative cannot see herself as *autonomous*; or that she cannot see herself or value herself as a rationally *reflective agent*, acting for reasons; or that she is committed to *practical solipsism*; or that she cannot coherently expect *other* people to respond to the reasons she addresses to them; or the like.[147]

It may be asked however, what is so bad about being unable to see oneself as autonomous or as rationally reflective and so on? Why should viewing oneself in these particular ways be of such paramount importance (and thus why does it entail reasons for action that are practically speaking, inescapable)? As Joyce writes "such things certainly have a nasty ring about them, but what does that ring really amount to?"[148] Joyce takes Christine Korsgaard to provide an answer: to ignore such reasons is to violate the "conceptions of ourselves that are most important to us… it is to lose your integrity, and so your identity,… it is to no longer be able to think of yourself under the description

[145] Richard Joyce, *The evolution of morality*, p. 195.
[146] *Ibid*., p. 195.
[147] David Copp, 'Moral naturalism and three grades of normativity', p. 35.
[148] Richard Joyce, *The evolution of morality*, p. 197.

under which you value yourself and find your life to be worth living… it is to be for all practical purposes dead or worse than dead"[149]

What then should we make of this argument? The charges Korsgaard makes certainly sound like they should be something we should care about. However, one immediate problem is that "these strong claims are supposed to reveal what is wrong about *any* moral transgression: failing to return a borrowed book, being rude to an undeserving waitress, pinching a morning newspaper from a hotel corridor."[150] But the fact that someone may occasionally do such things, and yet, on the whole, lead a happy, reflective, and satisfying life, should lead us to question whether this really is a good explanation of why we must care about such moral considerations. Thus, one potential weakness appears to be that the self-conception strategy cannot adequately account for the full range of moral transgressions. Presumably however the self-conception theorist is likely to respond that such minor moral transgressions do not perhaps cause you to be "dead or worse than dead" and that in these cases such rhetoric might not be warranted, but that they *do* create a kind of incoherence or some kind of feeling of loss of integrity, or that one is acting against one's character in some small way when one makes minor transgressions. Indeed, it makes sense that in such cases the strength or weight of the reasons in question are likely to be more minor to accompany the more minor transgressions. Joyce's complaint about the self-conception argument not accounting for very minor transgressions therefore is not a conclusive reply.

However, there is a deeper problem with the self-conception strategy. The problem is that the self-conception strategy simply assumes rationally deciding to ignore such reasons from time to time (and thus acting against one's self-conception) will result in wholesale rejection of continuing to be able to conceive of oneself in such a way. But why should we simply accept that the issue is settled by the self-conception theorist's claim that we *must* cease to view ourselves in this way if we only

[149] Christine. Korsgaard, *The sources of normativity*, p. 102.
[150] Richard Joyce, *The evolution of morality*, p. 197.

occasionally act in ways contrary to it? Simply claiming this is not enough; the authoritative normativity of moral claims cannot be ultimately dependent on the self-conception theorist's assertions that we 'must'. David Copp illustrates this point nicely using the story of the ring of Gyges from Plato's *Republic*:

> Gyges values the power and love he will achieve if he carries out his plot[151], and this means that, if he is fully rational and thinking clearly, and if Korsgaard's theory is correct, he must value his reflective agency. Moreover, if Korsgaard's analysis is correct, and if Gyges understands morality, he must then understand that carrying out his plot would conflict with his valuing his reflective agency. Yet he also values power and love, and he understands that carrying out his plot would help him achieve a life of power and love. So he might ask, "Why should I not carry out my plot? Why should I be moral?" For all that Korsgaard has shown, it seems to me, his asking these questions would not indicate either that he is not fully rational or that he does not understand morality. Nor need it indicate that he does not value his reflective agency.[152]

Just because Gyges values one way of seeing himself, it does not follow that he cannot also value the things he can attain by carrying out his plot. A person who values their own identity as an autonomous reason-responsive individual can still rationally wonder whether they should act against moral reasons, and sometimes decide to do so if there are other reasons that seem to have more weight. Ultimately, the claims of the self-conception theorist come down to the assertion that "moral reasons cannot be ignored, because if you do, something bad will happen to you or the way you think about yourself". But put in this way, such a claim hardly seems to be an adequate account of the inescapability and authority of morality that Joyce is trying to locate, especially when it is rationally debatable as to whether the "something bad will happen to you" in question can be cashed out in any substantial way that one cannot decide to ignore. The self-conception theorist would no doubt question this response: if we were really rational and clearheaded we *would* see that it is impossible

[151] Gyge's 'immoral' plot is as follows: "Having made his discovery [of a ring that could make him invisible] he managed to get himself included in the party that was to report to the king, and when he arrived he seduced the queen, and with her help attacked and murdered the king and seized the throne." Plato, *The Republic*, Part I, Book IV.

[152] David Copp, *Moral naturalism and three grades of normativity,* p. 37.

to ignore such reasons. But, the fact that many philosophers, who we would consider rational by any other standard fail to see this, and that there is little more that can be said to convince them (that is non-circular) means we *can* legitimately question this naturalization of the authority and inescapability of morality. This however is just one attempt to show that the authority of morality can be naturalized, and thus its failure would not show that the project is doomed.

In response to the failures to provide an adequate account, a supporter of moral naturalism with practical clout is likely to raise at this point what is called the 'partners in innocence' strategy (or alternatively 'companions in guilt' strategy – which term one opts for usually depends on the side of the debate one's support falls). The tactic here is to point to other kinds of normativity that have a feature that is analogous to practical clout – other kinds of 'ought-ness' of a particularly inescapable and authoritative kind – and to observe that there are not similar kinds of sceptical doubts about naturalising *these* kinds of normativity in other domains. Typical examples include logical normativity (facts about *to-be-deduced-ness*) or epistemological normativity (facts about *to-be-believed-ness.*) That is, few doubt whether there are facts about what one ought to deduce given a certain set of premises or about what one ought to believe given certain other beliefs.[153] These 'benign' cases force those who identify something wrong with practical clout to come up with a reasoned distinction between the moral case and the epistemological or logical cases of normativity, and if they cannot, this provides reason for accepting that the moral case is perhaps also unproblematic. Of course, at this point, it is open to the critic of naturalising practical clout to claim that these other kinds of normativity are also unacceptable and cannot be naturalised, but doing so is usually considered a counter-intuitive option.

What can be concluded from the 'partners in innocence' or 'companions in guilt' argument? Firstly, the argument does not actually deal with the difficulties of naturalising the various kinds of normativity – instead it points to different kinds of normativity in an attempt to shift the burden of

[153] Richard Joyce, 'Moral anti-realism'.

proof from those providing a naturalistic account of morality to those who criticise such accounts. Providing an account of how these other non-moral kinds of normativity can be naturalistically accounted for, how they reduce to or supervene upon natural facts, is unlikely to be straightforward: reductions of these kinds of abstract entities that depend upon human mental capacities are in general very hard to provide. How the mind or the 'mental' reduces to the phenomena that science deals with, is an unresolved problem in philosophy, and thus introducing other kinds of reductions that involve this same reduction in an attempt to clarify the moral case, is of questionable value. On the whole, pointing to other kinds of normativity does little to settle the debate either way. It does however show that morality is not alone in being a phenomenon that involves reductions that are difficult, and highlights that while there are not straight forward ways to philosophically account for or deal with such phenomena, this fact does not in general cause us to automatically jettison such kinds of normativity as errors.

Joyce takes himself to have "devoted quite a lot of energy to pursuing the possibility of a moral naturalist finding inescapable practical authority in the world via connecting it to a naturalistically respectable account of sufficient reasons"[154] and he concludes that "this avenue has led to a dead end."[155] There is more to be said about the success of this conclusion, but first it is worth considering the other response to Joyce's argument: that practical clout is not in fact a necessary element of a suitable moral naturalism.

3.2.6.2 Moral naturalism without practical clout

The second response to Joyce's argument against moral naturalism is to claim that we can have morality without thinking that it must provide us with authoritative inescapable "to-be-doneness" – that we can have morality without practical clout. An example is useful to show where the two

[154] Richard Joyce, *The evolution of morality,* p. 198.
[155] *Ibid*., p. 198.

positions diverge. Take for instance what is supposed to be a straightforward example of a self-evident moral obligation from Peter Singer:

> I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing.[156]

Joyce would argue that the concept of morality requires that a moral obligation to help the drowning child, is of a kind that provides a reason one cannot easily ignore. That is, the reason has some deliberative weight that one cannot ignore if one understands the situation. This does not mean that it is necessarily motivating but the reason for action here does not depend upon one's own desires nor anyone else's. It is simply *required* in some primitive, unavoidable sense. Joyce's hypothetical opponent will agree that there is a moral obligation to help here. And they will certainly also agree that almost anyone in this situation will have *strong* reasons to act, ones that will override worries about muddy clothing. But the moral-naturalist-without-clout will not accept that there is an unavoidable, desire-transcendent reason to act. Instead, according to them, there will simply be a range of contributing reasons for action; reasons that are based upon the everyday properties, relations, institutions, and so on, that are part of the natural world. Perhaps these reasons will be based on reflection on what kind of person one is, or thoughts about if oneself or one's child were in a similar situation (some kind of universalization of maxims or 'golden rule'), or considerations of the consequences of acting in one way or another, or alternatively perhaps reasons based on empathy and sympathy and attitudes towards avoidable suffering and loss and so on provide the justification. Whether the moral-naturalist-without-clout settles on some combination of the above, or perhaps different justifications not listed, the only difference (according to them) between their account of wrongness, and the kind of account Joyce would give, would be in the fact that they would deny that all these things amount to practical clout. The reasons for action will most likely be strong reasons, and if the individual in question is at all motivated by moral considerations, then they will have reason

[156] Peter Singer, 'Famine, affluence, and morality', p. 231.

to act. But there is nothing with any more *authority* than other kinds of reasons; the universe does not 'step in' and make the connection between the individual recognizing it is a moral obligation and the individual having a reason to act that they could not avoid having, no matter the contingent facts about them.

Paul Bloomfield has adopted this response, of denying that morality has practical clout, taking it to be obviously correct:

> What is of the last importance, and ignored by Joyce at what I saw as the culmination of the book's argument (p. 199-209), is that the ancient Greeks had recognizably moral systems, yet no analogous conception of 'practical clout'. Given Greek moral theory, we can see that practical clout is not a required feature of morality as Joyce, and many other philosophers, have suggested. Indeed, as G. E. M. Anscombe has famously argued, albeit in different terms, practical clout is a peculiarly modern feature of morality. Seen properly, morality without 'practical clout' is like combustion without phlogiston.[157]

David Copp also ultimately concludes that this kind of practical inescapability and authority is not a necessary feature of morality (Copp uses the term 'authoritative normativity' as a synonym for 'practical clout'):

> There is no definitive answer to the "Why be moral?" question of the kind that believers in *authoritative normativity* have in mind…There is no way to put all such doubts to rest in rational reflective persons. In my view, there is no answer to the "Why be moral?" question that lays it to rest by showing that it does not raise a serious practical question that could indicate an indecisiveness about morality in a rational person who had a clear understanding of morality. If this is correct, morality does not have *authoritative normativity*.[158]

This strategy is also advocated by Stephen Finlay in an exchange of four articles[159] with Richard Joyce. Finlay's argument is based around his view of morality whereby statements are "...oughts applying or

[157] Paul Bloomfield, 'Review: The evolution of morality', p. 179.
[158] Emphasis added. David Copp, 'Moral naturalism and three grades of normativity', pp. 40-41.
[159] Stephen Finlay, 'The error in the error theory', Richard Joyce, 'The error in "The error in error theory"', Stephen Finlay, 'Errors upon errors: A reply to Joyce', and Richard Joyce 'Enough with the errors!: A final reply to Finlay'.

ascribed to agents independently of their desires, but with normative authority that is relative to and contingent upon those agents' desires."[160] Finlay's position appears somewhat contradictory at first glance, for it sounds like it is trying to have its cake (the authority of oughts applying to agents independently of their desires) and eat it too (the authority of oughts being dependent only on that agent's desires – anything goes!). The question this immediately suggests is whether these oughts are the kind that morality actually deals in or is this an ad hoc characterisation of morality to avoid error theory. The answer to this question essentially what the debate between Joyce and Finlay is over, but it is worth noting that at first glance Finlay's position is highly counter-intuitive and seems to require an explanation of why its apparently contradictory claims are not in fact in conflict with each other.

So according to Finlay, he agrees with Joyce that practical clout[161] is not feature of morality that could exist – that 'objectively valid' moral imperatives that describe un-relativized normative facts not contingent on or relative to any agent's desires, institutional subscriptions, or ends, are impossible. However Finlay does not think this is problematic, he adds his voice to the camp of those who argue for moral naturalism without practical clout being unproblematic: "The bottom line is that they [Joyce and Mackie] and I hold basically the same view about what kinds of things the world contains; what we disagree about is whether this is enough to make some moral claims true."[162]

Joyce is aware that this point about the necessity of practical clout is contentious. He submits that there seems to be no established way of settling such a dispute where people's intuitions diverge[163, 164] and agrees with Michael Smith, David Lewis, and Mark Johnston who claim that which way one goes on the issue can come down to one's temperament:

> Strictly speaking…genuine values would have to meet an impossible condition, so it is an error to think that there are any. Loosely speaking, the name may go to a

[160] Stephen Finlay, 'Errors upon errors: A reply to Joyce', p. 541.
[161] Although Finlay prefers to use his own terminology rather than Joyce's suggested 'practical clout' despite his paper being an explicit response to Joyce's and John Mackie's arguments for error theory.
[162] *Ibid.,* p. 541.
[163] Richard Joyce, *The evolution of morality*, p. 200.
[164] Richard Joyce, 'Enough with the errors! A final reply to Finlay', p. 10

> claimant that deserves it imperfectly…What to make of the situation is mainly a matter of temperament.[165]

However, Joyce thinks that we can do better than this – that we can go beyond blunt charges of asserting or denying that practical clout is a necessary feature of morality. He argues that by looking at the way morality is used we can assess whether some feature of the concept is necessary. If morality did not have practical clout, Joyce thinks we would be unable to use morality as we do. As he puts it, if we look at "what the concept is *used for*, what practices it undergirds, and then ask whether a revised concept, with the problematic element discarded, could carry on playing that role",[166] then we will be able to assess whether this discarded element was in fact a necessary component.

Joyce thinks that if morality does not have practical clout, then moral reasons will not always provide us with reasons to act (that is they are either escapable, or sometimes do not have any authority), and that morality would therefore be only contingently connected to what we have reason to do.[167] He argues, if it is only contingently true that moral reasons result in actually having a reason to act, then there would be cases where we can say someone acted morally wrongly, for example by committing a murder, but that they had no reason to do other than they did. And this he thinks "imparts an extremely odd flavour to morality."[168] He thinks that we would not be comfortable asserting both that the murderer's actions are wrong, and then in the next breath claiming that they had no reason at all to refrain from committing the murder. Joyce thinks this makes morality odd because it reduces our ability to morally criticize wrongdoers – why should they care what we have to say, if even *we* admit that they had no reason to act other than they did? And if our ability to criticize wrongdoers is diminished, morality would be unable to be used as it is: it would have insufficient 'oomph' to provide normative guidance to everyone's lives.

---

[165] Michael Smith, David Lewis, Mark Johnston 'Dispositional theories of value', p. 93.
[166] Richard Joyce, *The evolution of morality*, p. 201.
[167] *Ibid*., p. 208.
[168] *ibid*., p. 204.

Finlay identifies the features of morality in Joyce's argument that show practical clout is necessary for it to function as follows[169]:

> *Address*: Moral imperatives or judgments are *addressed* as categorically applicable statements, to audiences who do not share the same concerns as the speaker. They are not addressed as statements contingent on the audience already subscribing to the same end as the speaker. This non-relativized usage is taken as evidence that moral judgments are not intended as relational in the way Finlay thinks morality is.[170]
>
> *Expectation*: Moral imperatives are addressed to non-subscribers with the *expectation* that doing so might influence the audience by provision of reasons they will realise or recognise the authority of.[171]
>
> *Disputation*: The fact that moral rightness is argued, even where parties appear to disagree about fundamental moral values, suggests there is an assumption of an absolute, non-relational moral truth to the matter.[172]
>
> *Reactive attitude*: our responses to those we believe have acted in morally significant ways are not contingent on what the actors themselves believe or the ends, desires, or normative standards they themselves subscribe to. This is taken as evidence that we expect moral norms to have practical clout – they would not serve their purpose if they did not.[173]

Finlay's response to these lines of argument is quite wide-ranging, but fundamentally rests on his positive theory of how morality is based on non-spoken institutional categorical imperatives. In support of this he also argues that moral conversations between parties that do not already share ends or normative standards are much less common than is assumed in metaethical discussion. He posits that most moral discourse within a society takes place between parties that share fundamental values and also that parties assume each other shares the same values; conversations with parties who do not share values, such as discussions with murderers or neo-Nazis are rare exceptions not the

[169] Finlay also identifies a number of other functions than are presented here but does not think they are serious candidates, so they have been omitted.
[170] Stephen Finlay, 'The error in the error theory', pp. 11-12.
[171] *Ibid*., pp. 11-12.
[172] *Ibid*., pp. 11-12.
[173] *Ibid*., p. 16-17.

norm. Joyce however argues this misses the point and that encounters with different fundamental moral values and outlooks are common place even if explicit conversations with parties that actually espouse them are not:

> Movies and novels are full of nihilistic or sociopathic baddies; the evil step-parent is a stock character of our children's fairy tales; and even small homogenous societies have their myths and religions that are full of destructive characters standing outside the accepted moral order…It is a challenge (to say the least) for Finlay's proposal to explain someone deliberating in her own mind over a moral dilemma or struggling with temptation toward immorality."[174]

Ultimately, Finlay's argument relies upon his theory that morality is a kind of shared institution, but one that everyone's subscription to is implicit. This is what allows him to argue that the authority morality has, is limited and dependent upon subscription to the institution of morality. For Finlay the authority moral speakers have (or at least the appearance of authority) is bolstered by them not being explicit about this relativistic component – the ends or goals that moral language assumes – and it is only because these are not made explicit that moral language has the surface appearance of dealing with fundamental moral disputes.

There is considerable detail to this discussion between Finlay and Joyce that has not been discussed but it is outside the scope of what can be covered here.[175] In the final analysis however, Joyce is not swayed by this argument, and I think rightly so. His view of Finlay's relativistic picture of morality is that it cannot do the job we expect morality to do – it lacks the practical weight and 'oomph' required to do its job.[176] As Joyce puts it "we want to do things with our concept of *moral rightness* for which the relativistic substitute just doesn't seem to provide license."[177] Thus, Joyce's argument for practical clout, that we could not use morality in the way we do if it did not have its distinctive form of authority

[174] Richard Joyce, 'Enough with the errors! A final reply to Finlay', pp 8-9.
[175] Being able to properly evaluate Finlay's position is also hampered due to a lack of a complete statement of his positive theory of how morality can be based on relativized imperatives despite this account being fundamental to his argument with Joyce.
[176] *Ibid.,* pp. 12-13.
[177] *Ibid*., p. 13.

is not thwarted by the considerations Finlay proposes. However, Joyce's proposals have also not persuaded all the cynics about practical clout either.[178] It appears that there is no agreed upon answer as to the question of whether morality has the feature Joyce calls practical clout.

While Joyce considers it a necessary component of morality, there are others who do not, and this divergence of opinion appears to be a problem for Joyce's argument against moral naturalism. As Joyce himself has noted, claiming that some feature "should count as a 'non-negotiable component' of morality has a tendency to lead quickly to impasse, for there is no accepted methodology for deciding when a discourse is 'centrally committed' to a given thesis."[179] It may be that there is simply be no matter of fact about practical clout if the concepts themselves are not determinate enough for such a conclusion to be established one way or the other, and different individuals and groups use the concepts in varying ways.

3.2.7 Does Joyce's debunking argument succeed?

Joyce has argued that morality is the result of our particular evolutionary history. If not for this particular genealogy, we would not have morality: we would not think in moral terms or use moral concepts at all, and therefore we would not have moral beliefs involving these concepts. Thus, if our moral beliefs are to be about moral facts, they must in some way be based upon (reduce to or supervene on etc.) this genealogy or its products. The prospects for such a naturalisation however, are not good according to Joyce. He attempts to show that morality cannot easily be accommodated as part of the usual natural world by showing that practical clout, a purported feature of morality, cannot be naturalised in an acceptable way. He discusses a number of possible ways that practical clout could be accounted for, but concludes that none of these are suitable – none provide the kind

[178] As Joyce himself readily admits, citing many including Gilbert Harman, Peter Railton, David Lewis, Mark Johnston, Jamie Dreier, David Copp, Jesse Prinz, Simon Kirchin, and Caroline West. See 'Enough with the errors! A final reply to Finlay', p. 12 and Richard Joyce, Simon Kirchin (eds.), *A world without values: Essays on John Mackie's moral error theory*.

[179] Richard Joyce, 'Moral anti-realism', sec. 4.

of inescapability or strength of authority that he thinks is necessary, in a naturalistically acceptable way. This is not the final word on Joyce's proposed error theory[180] but so far, the presented considerations are not successful in refuting Joyce, which makes his argument an interesting case study for the present thesis.

3.2.8 Implications of error theory

In the final chapter of *The Evolution of Morality* Joyce discusses the potential practical implications of his position. Joyce thinks that it is far from clear what implications his version of moral scepticism would have if it did in fact turn out to be correct. His discussion is interesting as he argues that we still have good reasons to act in accord with what morality recommends, and the considerations he advances in support of this, appear to be similar to the considerations put forward by many moral naturalists. Joyce insists that the accusation that "anything goes" follows from his kind of moral scepticism is unfounded. He writes "Moral scepticism amounts to the recognition that there is, or may be, nothing distinctively *morally* wrong with stealing, but it is absolutely not to be identified with the proposal that ordinary people have no reason at all to refrain from stealing – and anyone who made such a jump would be committing a grave mistake."[181] So, if according to Joyce, moral reasons do not matter, what kind of reasons *do* we have for thinking we should not steal? Why would it be such a "grave mistake" to reach the conclusion that stealing would be ok, if we do not need to worry about morality?

Joyce mentions a couple of answers although his discussion is brief. One reason Joyce gives is that moral emotions and sentiments are likely to persist, despite any epistemic ban on moral beliefs. Such

[180] Joyce himself accepts this, pointing out this is not the only argument or consideration weighing on the issue: "I should like to reiterate how narrow-minded it is to think that the moral error theory stands or falls entirely on this...Finlay goes so far as to say that he considers 'morality provisionally vindicated if Mackie's and Joyce's arguments are refuted'...The above argument could be totally unsound (and quite possibly is) and the error theory could still be persuasive." From Richard Joyce, 'Enough with the errors! A final reply to Finlay', p. 2.
[181] Richard Joyce, *The evolution of morality*, p. 224.

moral emotions he thinks play important instrumental roles in our social interactions. For example, the emotion Joyce calls 'indignant anger' – an emotional response to those that cheat or free ride or 'rip off' others – serves the purpose of discouraging such behaviour in others through giving motivation to punish, exclude, or extract reparation from transgressors, despite doing so being costly.[182] Even if such a response is costly in the short term, establishing a reputation that ensures others know you will respond in this way can result in long-term gains by deterring others from cheating, and signalling to others that you are a trustworthy co-operator – resulting in profitable on-going relationships. Thus, Joyce thinks that not only are moral emotions likely to persist, but that we *should* permit them to continue to play a motivational role in our lives since doing so produces desirable outcomes (he claims the "should"[183] here is one of prudence). Joyce even goes as far as claiming that a person will be "practically *irrational*"[184] if they do not allow moral emotions to play a role in their practical deliberations.

The persistence of moral emotions is not however the main reason Joyce thinks we should continue to refrain from stealing. He thinks that "An ordinarily situated person has many reasons to refrain from stealing – robust and plain reasons"[185] and that "The basis of some of those reasons is the fact that for social creatures, as humans are designed to be, major and irreplaceable satisfactions are to be had from sincere participation in a community."[186] His discussion suggests that people have prudential reasons (they will benefit from acting morally) and also that we have desires and needs that cannot be met in any other way than by acting morally - there are "major and irreplaceable satisfactions [that] are to be had from sincere participation in a community." What does this phrase about "major and irreplaceable satisfactions" mean? Presumably Joyce means that acting in accord with morality (by not harming others, not stealing, reciprocating when appropriate, cooperating fairly

[182] *ibid*., chapter 4.2 and 4.3, pp. 108-123.
[183] *ibid*., p. 228.
[184] *ibid*., pp. 227-228.
[185] *ibid*., p. 224.
[186] *ibid*., p. 224.

and so on) is in general a pre-requisite for acceptance in a community and for obtaining the range of goods that such acceptance makes available. Goods such as friendship, respect, trust, participation in all kinds of social activities and cooperation, freedom from guilt and shame, and the freedom from distrust and dislike of others are all available only by abiding by the norms that are present in communities. These norms provide assurance to members of the community that other individuals are suitable partners for these interactions which result in such goods. Joyce presumably thinks that these goods are almost universally desirable or preferable for social creatures like us, and thus the vast majority of people have reasons to pursue them. These considerations appear to amount to the fact that everyone has reasons to act in accord with and support morality.

There is little to find fault with in these considerations about the social nature of humans. The difficulty with Joyce's discussion is that these considerations for continuing to act in accord with what morality recommends, appear to be similar to the kinds of considerations put forward by the moral naturalist who thinks we can have morality without practical clout. If these reasons are sufficient for supporting his claim that we should act in ways that are in accord with what morality dictates, then Joyce has put forward a (somewhat unelaborated) version of moral naturalism without practical clout. No doubt Joyce would claim that such a conception of morality does not do justice to the usual concept of morality used in contemporary philosophy (not that there is likely to be a consensus on such a concept). So, on the one hand Joyce has identified what he thinks is an error in the commitments of moral discourse, but on the other, he also agrees with moral naturalists (of the without practical clout variety) that there are good reasons for acting morally and thus we have good reasons for maintaining moral practice as it is.

Does this amount to morality being debunked? The answer that I suggest is that in the sense that it forces us to revise our views of the nature of morality it might, but in perhaps a more important practical sense it does not. Consider the following analogy. Imagine that there is a primitive tribe of people. This tribe plants willow trees around the area in which they bury their deceased and believe

that the spirits of their ancestors infuse the willow trees with special healing powers. Because of this, they believe the magically infused bark and leaves of these trees can be used as a medicine that effectively reduces pain and fever, which in fact the bark does. Suppose then that this tribe does eventually come in contact with the modern world, and they learn that it is in fact salicylic acid in the willow bark that is responsible for the analgesic and antipyretic effects, and not the spirits of their ancestors causing these effects. Is their belief in their remedy for pain and fever debunked? In one sense yes: the causal mechanism they thought was responsible for the beneficial effects is not in fact what caused the beneficial effects (their ancestor's spirits are not playing any causal role that they believed they were). But in another clear sense their belief is not debunked: the beneficial effects of the remedy were real, and will continue to be, even though the element they thought was necessary (the spirits) for the medicine is undermined.

Similarly, if some or all of moral discourse has a commitment to a feature that is in fact mistaken, then in one sense it has been debunked. This is what Joyce's argument, if it is sound would establish. He would have demonstrated that when people make moral claims that things are right or wrong, that at least part of what they are doing is making a claim about a concept that is incorrect: they are claiming that the imperatives they put forward have a kind of super-natural authority and inescapability that they do not in fact have. But in another more practical sense morality has not been debunked: Joyce's argument has not shown that we should cease to act in accord with what morality recommends. Indeed, Joyce himself thinks we still have good reasons for acting morally. Therefore, just as the fictional tribe's practice of using the bark and leaves of willow trees was still an effective remedy even after their beliefs were shown to contain factual errors, so too morality still works for us, regardless of the fact our moral discourse contains a kind of factual error. It still works in the sense that it functions as it previously did to provide social cohesion, discourage actions that negatively impact on others, makes possible cooperative and productive interactions, and in general facilitates our living as the ultra-social creatures that humans are.

Thus, the conclusion is this: if by "debunked" we mean that we should abandon our concern for acting morally, then Joyce's argument does not show that morality has been debunked – the evolutionary explanation of morality does not 'explain away' our reasons for acting on what morality 'recommends'. If by "debunked" we mean Joyce's argument would show that moral discourse involves a kind of factual error – that moral discourse is committed to imperatives that purport to have a kind of authority and inescapability that they do not in fact have – then we can accept that Joyce's argument may 'debunk' this apparent factual appearance of morality. Whether Joyce's argument for this conclusion is sound however, depends on whether morality has a "non-negotiable commitment" to practical clout. And this, it seems, is hard to establish, as Joyce's argument is not conclusive on the matter and many disagree with him that it is a necessary feature. Conclusions such as Joyce's, are among the range of considerations used to support various 'non-factualist' meta-ethical positions including moral constructivism (the idea that moral facts exist only insofar as we decide what is morally right or wrong), non-cognitivism and its modern derivatives such as expressivism and Simon Blackburn's quasi-realism, fictionalism, and a range of subjectivist and relativist positions. The arguments for such meta-ethical positions are almost *all* controversial (in the sense that consensus is at best divided if there exists any at consensus at all) and thus it should be of little surprise that the same is true for opinions on Joyce's argument. Nevertheless, Joyce's argument is a good model for how an empirically based consideration could have significant implications for meta-ethics and our understanding of moral truth, and I will discuss these implications further in chapter 4.

3.3 Sharon Street's Darwinian dilemma

As Joyce's argument shows the causal origins of one's beliefs have the potential to make us re-evaluate the epistemic credentials of those beliefs. Whether a particular belief's genealogy is of the undermining or reinforcing variety is a question of what has happened in the past and whether that past is conducive to the beliefs that were formed being true. Sharon Street has made a number of

contributions in recent years that also advance arguments about the potentially undermining role that evolution can play as part of the causal origin of our moral beliefs[187]. Street argues that if the evolutionary genealogy of morality (as outlined in chapter 1) is correct, it poses a dilemma for the moral realist that forces them to choose between endorsing an unscientific theory and abandoning at least some of their realist aspirations.

Below, I outline the target of Street's argument, moral realism, and then describe how the dilemma supposedly forces the realist to choose between moral scepticism and an improbable and scientifically indefeasible theory of the evolution of human sociality, cooperation, and morality. After this I briefly describe the strategies the realist may wish to take up in an attempt to avoid the Darwinian dilemma and then examine some of the responses that have been made to evolutionary error theories that are applicable to both Street and Joyce's arguments.

3.3.1 Realist theories of value

The targets of Sharon Street's Darwinian Dilemma are what she calls 'realist theories of value'. The defining claim of realist theories of value is that "there are evaluative facts or truths that hold independently of all of our evaluative attitudes."[188] The constituents of this definition are broken down as follows.

By 'evaluative truths or facts' Street is talking about propositions such as 'X is a reason to Y', 'one should or ought to X', or 'X is good, valuable, worthwhile, or morally right'. These are the moral truths or facts that a realist theory of value deals in.

By 'evaluative attitudes' Street means the relation we have to the 'evaluative truths or facts'; the judgments we make about them. These 'evaluative attitudes' include conative and cognitive states

[187] Sharon Street, 'A Darwinian dilemma for realist theories of value', 'Reply to Copp: Naturalism, normativity, and the varieties of realism worth worrying about', and 'Coming to terms with contingency: Humean constructivism about practical reason'.
[188] Sharon Street, 'A Darwinian dilemma for realist theories of value', p. 110.

such as desires, attitudes of approval or disapproval, judgments about what one ought to do, what one has a reason to do, and what is morally right or correct. They include morally significant emotions and sentiments; attitudes that favour cooperation, reciprocation, and fairness.

By 'hold independently' Street means that there are evaluative facts that could be true independently of the entire set of actual evaluative judgments, attitudes, or beliefs we hold, have held, or will hold. An example Street gives of a view that includes this kind of independence from our evaluative attitudes is what Russ Shafer-Landau has called 'Stance-independent'[189] realism.

Consider the statement 'Hitler was morally depraved'. According to a realist about value, it might be (and probably is if our moral beliefs are similar to those of most moral philosophers) true that Hitler was morally depraved and this truth will hold independently of any stance (evaluative attitude) that we, Hitler, or indeed anyone else, might take toward that truth. The fact that Hitler was depraved does not rely on anyone actually holding the attitude or view that he was depraved. It is possible on this view that everyone that exists actually thinks Hitler is admirable, yet according to the realist it would still be true that Hitler was depraved (and everyone else had false beliefs).

3.3.2 Evaluative attitudes saturated by Darwinian influence

The Darwinian Dilemma is the result of the recognition that evolutionary forces have had a tremendous role in shaping our evaluative attitudes and beliefs. It is worth elaborating exactly what Street means by "a tremendous role in shaping" as the strength of these influences is often cited as a point of weakness by opponents of the Darwinian dilemma.[190] The potential range of evaluative views that we could potentially hold is very large (possibly infinite). The size of the subset of these evaluative

[189] Schafer-Landau 2005, *Moral Realism: A Defence.*
[190] For example, David Copp presses this point in 'Darwinian skepticism about moral realism', pp. 190, 203-204.

views that humans actually hold is by comparison very small. Street gives some examples of widely held evaluative views that might be similar to the following (pp. 115-116):

a. The fact that something would harm oneself is a reason not to do it.
b. The fact that someone helped you is a reason to help them in return, or a reason to praise and thank them.
c. The fact that someone has cheated or taken advantage of a cooperative situation is reason to stop cooperating with them, condemn them, and perhaps punish them.

These sorts of prototypical evaluative attitudes are widely held. Further, evolution can do a very good job of explaining and even predicting further attitudes of this kind. Compare a) to c) with some other potential evaluative beliefs that humans could potentially hold (Street 2013, pp 116):

a'. The fact that something would be harmful to oneself is a reason to do it.
b'. The fact that someone has helped you is a reason not to help them, and is a reason to condemn and disparage such behaviour.
c'. The fact that someone has cheated or taken advantage of a cooperative situation is a reason to cooperate further with them, praise them, and perhaps reward them.

But these inverted versions of a) to c) still only covers a tiny part of the space of possible evaluative views. Consider some other potential evaluative beliefs:

d'. One should be singularly occupied with the goal of obtaining as many pineapples as possible.
e'. The good life consists in climbing the highest point visible at all times and diving off of it head first.

Evolutionary biology provides an explanation of why tendencies such as a) to c) would have been more conducive to reproductive success whereas tendencies a') to e') would have had clearly negative or deleterious effects.

There is an important clarification that is necessary to this story; the connection between the content of our evaluative beliefs and the forces of natural selection that influenced them is an *indirect* one. Evolution works on traits that are genetically heritable. It is unreasonable to think that full-fledged, linguistically infused evaluative judgments could be genetically heritable. The evaluative judgment that "She owes me something in return" for example, is an implausible candidate for natural selection to work on.

Instead, it is what Street calls our "basic evaluative tendencies"[191] that were genetically heritable traits that natural selection worked on. A basic evaluative tendency is an "unreflective, non-linguistic, motivational tendency to experience something as 'called for' or 'demanded' in itself, or to experience one thing as 'calling for' or 'counting in favor of' something else".[192] It is these more basic evaluative tendencies that it is plausible to think were genetically heritable and due to genetic differences, that have been at play over the course of much of our evolutionary history. There is a striking continuity that is observable between the types and kinds of evaluative judgments we make and the basic social tendencies observable in ethology of primates and other animals. These basic evaluative tendencies matter however, as they are the basis of later, more sophisticated 'fully-fledged' evaluative tendencies.

To see why this is so, consider a counterfactual case: imagine the early evolutionary history of humans took a path similar to that of eusocial insects such as Ants or Bees, or perhaps that of a species of one of the solitary *Felids*. If this were the case, we would expect our basic evaluative tendencies to be very different in a number of ways. In the latter case we would expect instincts and evaluative tendencies that produced behaviour that was territorial, confrontational to conspecifics except in mating contexts, and had little to no altruistic behaviours towards non-kin or distantly related kin. In the former case we might find very different evaluative tendencies that favoured extreme sociality

[191] Sharon Street, 'A Darwinian dilemma for realist theories of value', p. 119.
[192] *Ibid.*

towards kin, strong social roles or hierarchy within families, and essentially belligerent or avoidant behaviour towards other conspecifics.

With these differences in the basic evaluative tendencies we would expect that, *mutatis mutandis*, our later more full-fledged (and counterfactual) evaluative tendencies would be very different from our actual full-fledged evaluative tendencies that follow from our actual basic evaluative tendencies.[193]

Thus, despite the indirect manner of influence of that basic evaluative tendencies have, they still exert a powerful influence on what our full-fledged evaluative tendencies end up being. Full-fledged forms of evaluative tendencies may have been a relatively late evolutionary add-on, but the basic evaluative tendencies still determined the direction and nature of the overlaid cognitive and more complex forms of evaluative judgment.

3.3.3 The Darwinian dilemma

Summarising the above background, the dilemma can be brought into focus as follows: normative beliefs could potentially be anything (conceptually at least, there is an enormous possibility space). The actual normative beliefs we hold are in fact a very narrowly constrained subset of the possibility space. Evolution provides a powerful explanation why these beliefs are constrained to this subset. Given these facts, we can enquire from the normative realist, why would our normative attitudes that have been shaped by evolutionary forces know anything of the mind independent normative truths posited by the realist? What, if any, relation is there between the mind independent normative truths and the evolutionary influences that actually gave us our particular moral psychology?

[193] Of course, without strong selective pressure for group living, cooperation, and sociality, it is likely language and other necessary elements for morality would not have come into existence at all.

Street argues that the realist faced with this question must choose one of two options and that neither option is likely to appeal to them. As choosing either of the options that follow from being a realist result in untenable positions when viewed through the Darwinian Dilemma, the conclusion Street recommends is that the realist should revise their starting point and reject mind-independent normative truth.

The dilemma the moral realist faces, is as follows. Either there is a relation between independent evaluative truth and the evolutionary forces that shaped our evaluative attitudes/beliefs about evaluative truth or there is no such a relationship. The first branching of the root of the diagram below (labelled as 'Darwinian Dilemma') reflects these two choices open to the moral realist: deny there is a relationship between "independent evaluative truth" and the evolutionary influences that shaped our psychology (C. in the diagram), or claim that there is some kind of relation between what happened in the evolutionary past and independent evaluative truth (B. below):

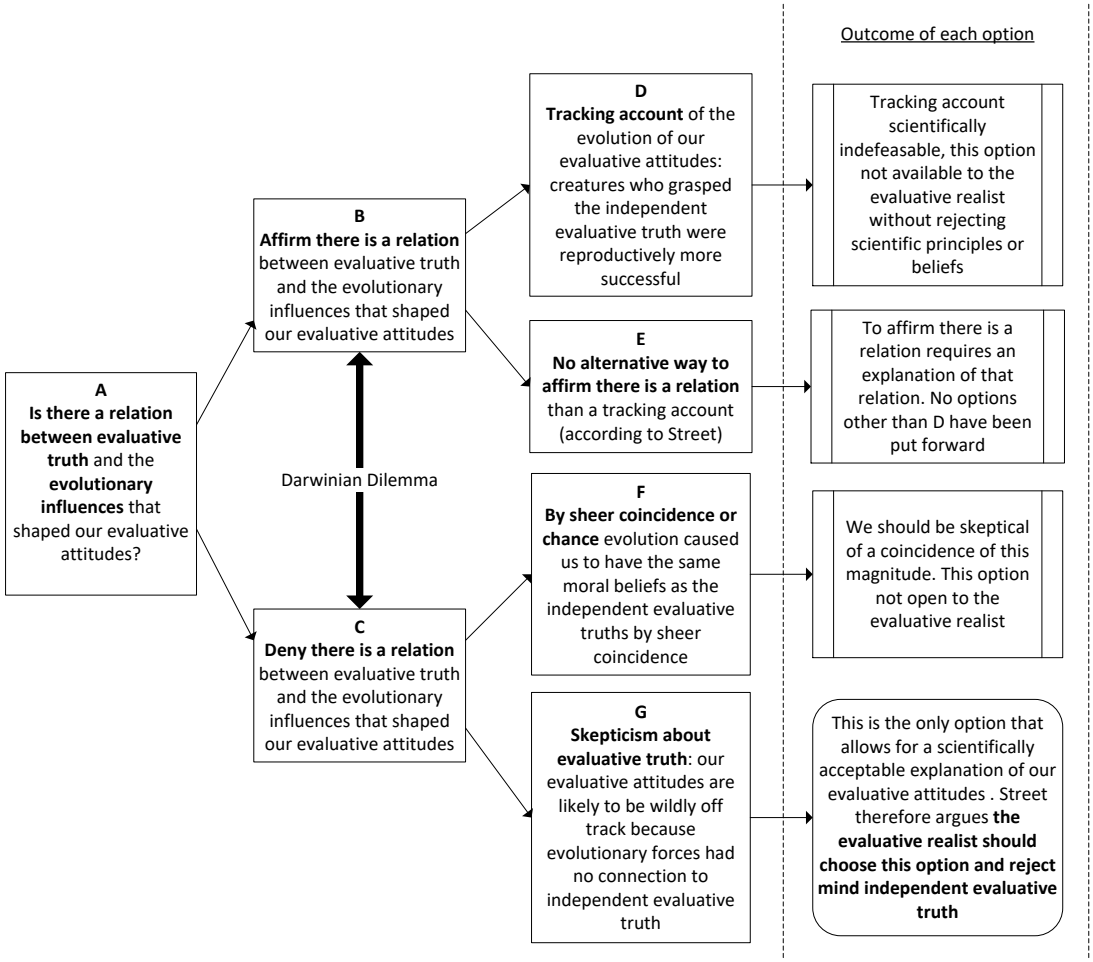


Figure 8: Sharon Street's Darwinian dilemma

If the normative realist opts to deny there is any relation (C), they face a further choice. They can choose to accept that because there is no relation between the evolutionary past and independent evaluative truth, whatever moral psychology evolved is overwhelmingly likely to produce beliefs that are not the same as the independent evaluative truths (G). The beliefs produced by the evolved moral psychology is overwhelmingly unlikely to overlap meaningfully with the independent evaluative truth because the set of potential evaluative beliefs (recall this includes even patently nonsensical ones such as 'one must obtain as many pineapples as possible') is so astonishingly large compared to the set of beliefs that our actual moral psychology produces. Street argues the result of this is a kind of moral scepticism, if evaluative truth truly is independent of the evolution of our moral psychology, then we should expect their "predictions" to be independent as well.

If the normative realist chooses option (F.) they are committing themselves to an astonishing co-incidence where despite the enormous possibility space, by sheer luck our evaluative beliefs have ended up being identical to the independent evaluative truths. Thus, Street thinks if they choose to deny there is a relation between independent evaluative truth and the evolutionary influences that shaped our evaluative attitudes, then the realist's only plausible option is (G.): evaluative skepticism.

Alternatively, if the realist chooses the other horn of the Darwinian dilemma and claims that there is a relation between the evolutionary past and our evaluative attitudes (B.), then they are committing themselves to a so called "truth-tracking account", where in the environment of evolutionary adaptation it promoted reproductive success to directly grasp ('track') independent evaluative truth (D). According to a tracking account, our ability to recognize evaluative truths in the evolutionary past conferred upon us certain advantages that resulted in differential reproductive success.

Street notes that this is an explanatory account that will be tractable to evaluation by the standards of science.[194] This explanation offers a hypothesis as to how some of the course of human natural selection proceeded and how specific features of human psychology evolved. In particular, our disposition to make certain kinds of judgments and believe certain kinds of beliefs is explained by the fact that we could somehow recognize the independent evaluative truth. The ability to discern these truths in turn proved advantageous for survival and reproduction.

A concrete example may be advantageous to understanding here. If we ask why a widespread tendency for humans to be good at detecting cheating behaviour exists (behaviour aimed at gaining the benefits of cooperation without paying the price), the explanation the tracking account gives is that it is true that it is wrong to cheat, and that it promoted reproductive success to be able to grasp (track) this evaluative truth. How exactly this tracking connection functioned is left unspecified, as

[194] *Ibid.,* p. 126.

there does not seem to be a plausible way for tracking to operate if the evaluative truths were truly independent.

This makes the difficulty of endorsing a truth-tracking account evident: there is an alternative hypothesis which is eminently more plausible and scientifically defeasible. The alternative hypothesis is that our tendencies to make certain kinds of evaluative judgments in the evolutionary environment of adaptation created a "link" between the circumstances our ancestors found themselves in and certain adaptive responses to the problems of social living. By making those evaluative judgments which they did, our ancestors were caused to respond by judging, feeling, and acting in ways that were advantageous for survival and reproduction. Street terms this the "adaptive-link account" because instead of tracking independent evaluative truths, our ancestors simply made judgments and formed beliefs that are "adaptively linked" to certain responses and beliefs.[195]

In the cheating example, the explanation the adaptive-link account gives is that it promoted reproductive success to hold certain beliefs about and attitudes towards, those discovered cheating. These beliefs or attitudes, for example, thinking that the cheaters' actions count as a reason in favour of punishing or excluding them, form the link between the judgment about the cheater's unfair actions and the adaptive response of penalising or shunning them.

Street argues there are at least three grounds on which the adaptive-link account is superior to the truth-tracking account. Firstly, the adaptive-link account is more parsimonious – it involves fewer entities and is simpler. The tracking account requires an extra element – independent evaluative truth – to explain why we make the judgments that we do. The adaptive-link account does not require us to posit this extra element, and thus can explain the existence of our evaluative beliefs and attitudes more simply, without needing the extra element of normative truth to play a role.

[195] *Ibid.,* p. 127.

Secondly, the adaptive-link account is superior to the truth tracking account because it is clearer. The mechanisms involved are easily identified and understood. The tracking account argues that we have the evaluative beliefs and attitudes we do because they were true and holding true beliefs about the independent evaluative truths was an evolutionary advantage. However, the truth-tracking account does not tell us *why* it was an evolutionary advantage to hold these beliefs. Why would those organisms that grasped such truths be reproductively more successful? Therefore, not only is the adaptive link account more parsimonious than the truth-tracking account, it is also clear in how it functioned, something that cannot be said for the truth-tracking account.

Thirdly, the adaptive-link account has superior explanatory power and tells us more about the explanandum in question. It can tell us why certain judgments are made rather than others, and show us the relation between our evaluative judgments – namely that they link the circumstances that engender evaluative judgments and the responses we make that would have been likely to promote the differential reproductive success of our ancestors. In contrast to the adaptive-link account, the truth tracking account has no explanatory or predictive power of this kind. It cannot tell us why our evaluative judgments are of the kinds and variety and subject matter that they are. When combined with the lack of an actual account of how truth-tracking account might latch on to truth and the comparative parsimony of the adaptive link account, Street thinks it is clear that the adaptive-link account is preferable over the truth tracking account.

Returning to the diagram, if the realist chooses to affirm there is a relation between our evolved moral psychology and independent evaluative truth (choosing the horn labelled B.), then they must support a truth tracking account. Unfortunately for the realist, a truth tracking account can be rejected on scientific and explanatory grounds as the adaptive link hypothesis is superior so this horn of the dilemma is an unattractive option overall for them.

The evaluative realist must therefore deny there is a connection between our evolved moral psychology and the independent evaluative truth (C.) and must choose between either (F) and claim

that there was a remarkable and improbable coincidence whereby what evolved just happened to overlap substantially with independent evaluative truths or they must accept that we should not trust our moral psychology to provide us with insight into evaluative truth – our moral beliefs are likely to be wildly off track (G).

3.3.4 If evaluative facts are identical to natural facts, can the dilemma be avoided?

Up in till this point, it may appear that the above argument is missing an obvious and important line of thought which would show the Darwinian Dilemma to be something of a false dichotomy. This line of thought represents a position Street terms the 'Value Naturalist'.[196] For the Value Naturalist, evaluative facts (including moral facts) are in some sense constituted by or identical with (some) natural facts. That is, they reduce to, supervene on, or in some other unspecified way, just are the natural facts in question about humans, their environment, and their "patterns of reaction to it" and therefore are the kinds of things that can play a causal role in explanations.

Value naturalism of this sort is often raised as a response to arguments that aim to show that realism is too metaphysically or ontologically strange or mysterious.[197] By locating moral facts as being identical to or the same as natural facts that figure in everyday explanations, there appears to be a way in which the evaluative realist can show that independent evaluative facts might play a straightforward role in the evolutionary story of our moral psychology. If there is a plausible account of why we might have evolved to track these particular natural facts, then there is an account available to the realist of the relation between evolutionary pressures on our evaluative judgments and the independent normative truth. If this is so it would diffuse Street's Darwinian Dilemma.

Street's response to this line of thought is to remind us of the commitments of the evaluative realist and show why such a value naturalist could not connect our evolved judgments with independent

[196] *Ibid.,* p. 112.
[197] As typified by John Mackie's 'Argument from queerness' in *Ethics: inventing right and wrong*, pp. 38-42.

evaluative truths in the manner suggested by the value naturalist and still remain a thorough-going evaluative realist. For Street, to be a genuine evaluative realist, one must hold that there are evaluative truths that will hold independently of our evaluative attitudes. Her target for the Darwinian Dilemma is any kind of normative realism that holds that there are evaluative truths that are entirely independent of our evaluative attitudes.

The following chart shows the structure of Street's response to the value naturalist. Starting from the left, the value naturalist argues that the evaluative facts *E* just are the same as some set of natural facts *N*. Street then asks for clarification: does the truth of *E* depend in some way upon our evaluative attitudes? If the evaluative facts do depend on our evaluative attitudes, then they are not sufficiently independent evaluative truths to count as a thorough-going realism that the Darwinian Dilemma targets. If this is the case, then her argument is successful in showing that mind independence cannot be a feature of morality.

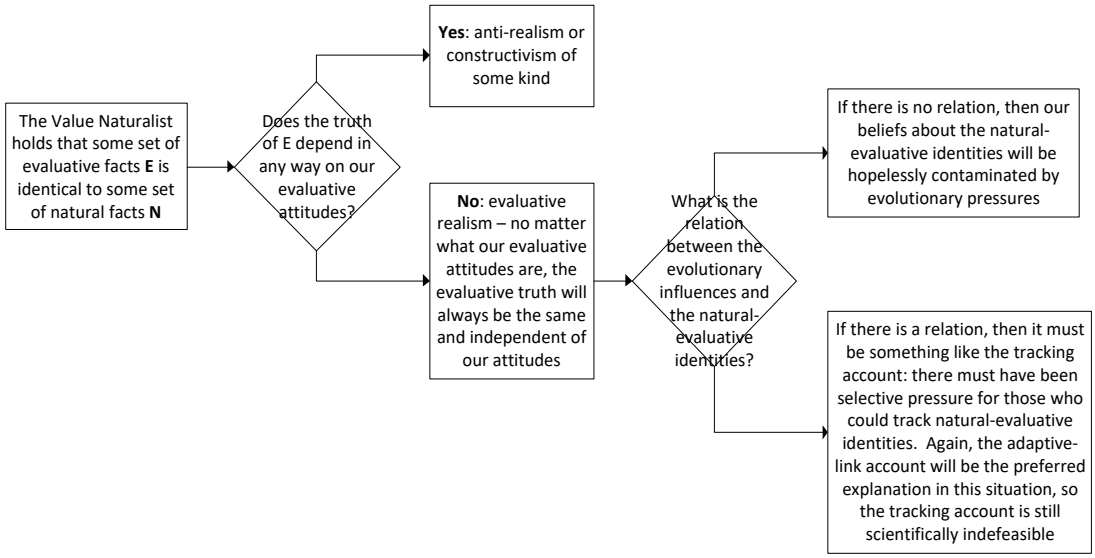


Figure 9: Street's response to the value naturalist

If on the other hand the truth of *E* does not depend in any way upon our evaluative attitudes, then this version of value naturalism can count as a "thorough-going" evaluative realism. However, this stipulation means that the resulting theory has other difficulties.

The evaluative beliefs and natural facts must be connected through the evolutionary process, but at the same time, they cannot be contingent on what our evaluative attitudes are, otherwise they no longer qualify as independent evaluative truths. This results in a serious problem for the evaluative naturalist, as they must then explain how the natural-evaluative identity is fixed. They must either deny there is a relation between the evolutionary influences and the natural-evaluative identities, in which case the natural-evaluative identities are thoroughly "contaminated" with the distorting influence of selective pressures. Or, if there is a relation, then it must be something like the tracking account: selective pressure for those who could track independent facts about natural-normative identities. But as explanations go, the adaptive-link account is clearly preferable. As Street concludes:

> To the extent that a view insists on there being evaluative facts which hold independently of all our evaluative attitudes, it is impossible to reconcile that view with a recognition of the role that Darwinian forces have played in shaping the content of our values. [198]

Once it is made explicit that our evaluative judgments are influenced in this way external forces, Street concludes that we are compelled to adjust our metaethical view to become some variety of antirealist.

3.4 General responses to evolutionary error theories

3.4.1 Capacity etiology versus content etiology debunking

William Fitzpatrick has criticised evolutionary debunking arguments, explicitly including those of Joyce and Street.[199] Fitzpatrick focuses on debunking arguments that aim to undermine ethics due to epistemic considerations. Specifically, he identifies what he takes to be two kinds of debunking arguments in the literature: those that focus on the etiology of our *capacities* for moral judgment and those that focus on the etiology of the *content* of our moral judgments. Fitzpatrick does not think either kind of debunking is ultimately successful: capacity etiology arguments do not raise any serious

[198] Sharon Street, 'A Darwinian dilemma for realist theories of value', p. 33.
[199] William Fitzpatrick, 'Debunking evolutionary debunking of ethical realism'.

problems for realism, and content etiology arguments rely on assumptions about the evolutionary genealogy and its relation to our moral practice which are not supported by the evidence.

Fitzpatrick proposes that capacity etiology arguments have roughly the following form. Our mental capacities, including mental capacities we employ in making moral judgments, are the product of evolutionary forces. These capacities were not designed to, and do not, track independent moral truths (even if they exist). Instead, evolution shaped our cognitive capacities to track facts "relevant to competitive gene propagation in ancestral environments."[200] Therefore, either the capacities we employ in making moral judgments do not result in cognition that reliably tracks moral truths, or, if they do, it would be by an implausibly lucky coincidence. Either of these options raise sceptical problems.[201] Fitzpatrick argues that this argument is invalid as it stands, but that it can be made valid by adding the following premise to the argument: the only way in which our cognitive capacities can result in judgments that are non-accidentally and reliably truth-tracking, is if evolution made them that way.

However, Fitzpatrick thinks this premise is false and the converse is likely true:

> This assumption overlooks the alternative possibility of our taking general cognitive capacities bequeathed by natural selection and developing them in cultural contexts, through relevant forms of training within traditions of inquiry into the subject matter in question, and thus making our cognitive dispositions in the relevant domain non-accidentally reliably truth-tracking.[202]

Applied to the moral case, this process can used to avoid the sceptical conclusion. It is enough that we take the capacities that evolution gave us (regardless of whether they originally produced off-track judgments with regard to moral truth) and learned to use critical reflection and reasoning in normative

[200] *Ibid*., p. 885.
[201] Adapted from *ibid.,* p. 884.
[202] *Ibid.,* p. 886.

contexts, and through this process we can learn to "reflect accurately on how it's good and right to live."[203]

Fitzpatrick maintains that because of the possibility of developing our moral capacities, we need not posit dubious stories about how our mental capacities might directly track moral truths or how through sheer luck they produce beliefs about independent moral facts. It is enough that we have developed the capacities evolution bequeathed us, and have then developed these through the use of critical reasoning, training, reflection, dialogue, experimentation, and so on to discover independent moral truths.

Fitzpatrick thinks proponents of the capacity etiology argument might accept that this story is true for some or perhaps even most cognitive capacities, that we can and do use cognitive capacities to produce beliefs and discover truths that go beyond what evolution designed those capacities for. However, Fitzpatrick claims that proponents of the capacity etiology argument would still reject this story as an explanation of *moral* capacities and independent moral truth. Their reply would be that moral reasoning is not an extension or development of any kind of reasoning that our moral capacities were designed for. It is instead a "*sui generis* domain of emotion-laden thinking that, if favored by natural selection, was favored not for accuracy but just for direct contributions to biological fitness."[204]

Fitzpatrick disagrees however, and thinks that moral reasoning can be viewed as an extension of forms of reasoning that our mental capacities *were* designed to do accurately. Fitzpatrick claims that both the logical reasoning and conceptual analysis required in the moral domain are continuous with our broader cognitive capacities. He cites David Enoch, who maintains that:

> Given a starting point of normative beliefs that are not too far-off, presumably some reasoning mechanisms (and perhaps some other mechanisms as well) can get us increasingly closer to the truth by eliminating inconsistencies, increasing

[203] *Ibid*., p. 887. Note that Fitzpatrick seems to think that it is obvious that being able to reflect 'accurately' on "how it's good and right to live" can be used synonymously with being able to track independent moral truth. However, it is far from obvious that we can conclude these two things are one and the same unless we already assume Fitzpatrick's argument is successful.
[204] *Ibid.,* p. 888.

> overall coherence, eliminating arbitrary distinctions, drawing analogies, ruling out initially justified beliefs whose justificatory status has been defeated later, etc.[205]

But crucially, this requires a starting point of normative beliefs that are "not too far-off"; beliefs that are already within reasoning distance of the independent moral truths. Firstly, it is not clear why, without it being an *ad hoc* response to the error theorist, this assumption should be granted. Note that it is not sufficient to think that the things that were helpful or harmful in social contexts of our evolutionary past seem relatively close in some respects to what we consider to be morally good and bad. The explanation required must show how, *even if those social contexts of our evolutionary past were entirely different,* we would still end up with the same things we considered morally good or bad.

Secondly, Enoch's argument starts with *beliefs* and not capacities that are 'not too far off'. At this point Fitzpatrick also slips into talking of conceptual *content* in his argument:

> At a formal level, we employ the same logical and analytic abilities in moral reasoning as in other forms of reasoning. And in terms of *conceptual content,* moral reflection and reasoning is continuous with broader evaluative and normative thinking that our cognitive capacities were plausibly designed to do accurately.[206]

It is not surprising that the conceptual content of morality is handled using the same kinds of reasoning and logic as other kinds of conceptual content – that is not what is at issue and does not contribute to showing that morality is not somehow *sui generis* with regard to truth-tracking. Fitzpatrick's argument does not explain how off-track moral capacities can get us to the *conceptual content* of the independent moral truth when there is no guarantee that the starting-point is anywhere near where it needs to be.

The above argument highlights a difficulty in Fitzpatrick's dichotomy of moral capacities versus moral content that results ultimately in his argument being unpersuasive. If he takes moral capacities to

[205] David Enoch, 'The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it', p. 428.
[206] William Fitzpatrick, 'Debunking evolutionary debunking of ethical realism', p. 888.

simply be the same kinds of reasoning and thinking capacities as humans use generally, then it is not clear that there is anything to them that marks them as being distinctively moral. It is no surprise then that on Fitzpatrick's account, the moral capacities are continuous with the general capacities that we *do* think can be truth-tracking. The sticking point is that his argument against moral capacities must choose some kind of distinctively moral content as a starting point for it to be a moral capacity that has an error-prone etiology. And as soon as this evolutionary starting point about content is introduced, Fitzpatrick needs to explain how this starting point relates to the independent moral truths (or in his version, how it is close enough to reason from that starting point to the independent truth) that does not depend on the contingent fact that they already seem quite close.

Because of this, it is unlikely an error theorist would be swayed by Fitzpatrick's capacity argument. However, despite this, they would likely agree with much of his conclusion about the evolutionary etiology of our moral capacities which:

> … impose[s] a constraint on realists as we go forward in developing a positive moral epistemology: any such account must at least square with our best scientific understanding of the sorts of capacities evolution gave us to work with, avoiding reliance on capacities we could not plausibly have developed from such psychological materials. That is a useful result and may pose an interesting challenge for some realists…[207]

For the present thesis, this is perhaps the more interesting result, and is closer to Joyce and Street's final positions about the nature of moral truth and moral realism.

There are a number of features of Fitzpatrick's arguments that there is not room to address, but perhaps the most interesting is that neither Joyce's nor Street's arguments can be cleanly categorised as either capacity or content etiology arguments. Instead, they are both best categorised as arguments about the etiology of both the capacity *and* content to greater or lesser degrees. Joyce and Street both take there to be something uniquely troubling about the epistemic origins of morality that includes

[207] *Ibid.,* p. 889.

both our moral capacities that are closely tied to our evolutionary past *and* the content of our moral beliefs that are too close to the content of the beliefs that are fitness enhancing in the proposed evolutionary genealogies of morality.

Fitzpatrick's content etiology argument depends mostly on arguing that the evolutionary influences on our moral beliefs are not as strong or as troubling as the error theorist proposes. He suggests that a debunking argument would only be successful if "evolutionary influence on our moral beliefs is somehow so distorting and difficult to expose that it undermines our ability to develop and employ reflective techniques to home in on independent moral truths."[208] The error theorist is unlikely to have any quarrel with this description as this is ultimately what they are themselves suggesting: that not only are the *capacities* we use to arrive at moral beliefs of a kind that are distorting but that the starting point of our beliefs, their *content*, is also thoroughly saturated with evolutionary influence.

3.4.2 The reliability of moral cognition

Benjamin Fraser argues that the evolutionary influences on our moral cognition and beliefs are in fact distorting, and, are actually more troublesome than evolutionary debunking arguments and their opposers have typically assumed.[209] Fraser focuses on the question of whether an evolutionary explanation of morality gives us reason to think that our moral faculty is, in actuality, unreliable. He argues there are a number of conditions that need to be assessed to establish whether an evolved capacity is epistemically reliable. The conditions he identifies are:

> 1. The *environment* condition: the mechanism is operating in an environment relevantly similar to that in which it evolved.
>
> 2. The *information* condition: information is not high cost, thus it is not adaptive to employ a cheap, error-prone mechanism.

[208] *Ibid.,* p. 901.
[209] Benjamin Fraser, 'Evolutionary debunking arguments and the reliability of moral cognition'.

3. The *error* condition: asymmetrical error costs are unlikely to have selected for systematic bias in the mechanism.

4. The *tracking* condition: the function of the mechanism is to track features of the agent's environment.[210]

Fraser does not think these features constitute necessary and sufficient conditions for epistemic reliability of an evolved capacity, but instead suggests that unless all or most of these are met, we should provisionally expect the evolved cognitive mechanism is now unreliable.

The environment condition highlights the need for cognitive mechanisms to be operating in circumstances which they are well suited to – usually this means the environment they were selected for or one that is relevantly similar. For example, 'fast and frugal' heuristics can have low costs for impressive results in the environment in which they evolved while providing disastrous results outside that environment.[211] While there is significant overlap between the evolutionary environment and today, Fraser concludes that the differences are sufficient to provide *prima facie* reason to suspect this condition is not met.[212]

For the information condition, the evolutionary benefit must outweigh the costs of building the cognitive mechanism if we are to expect it to be reliable. To do so, the evolutionary cost of accurate information must not be so high relative to the cost of errors that it would be more adaptive to adopt an error-prone but cheap mechanism. The kinds of costs that are relevant are the costs of gathering and processing relevant information. If the moral facts in question are non-natural facts, then it is not clear at all how to assess the costs of gathering and processing this information, so Fraser does not attempt to do so. Alternatively, if there are no moral facts, then no amount of information could contribute to epistemically reliable inputs. Focusing on the non-question begging option, that naturalistically respectable moral facts exist, Fraser suggests that the relevant kinds of information

[210] *Ibid.,* p. 461.
[211] P Todd, G Gigerenzer, 'Simple heuristics that make us smart'.
[212] Benjamin Fraser, 'Evolutionary debunking arguments and the reliability of moral cognition', p. 465.

will include "information about the intentions, motives and interests of other agents, and the consequences of actions, especially harms and benefits to others."[213] Fraser judges the costs of gathering this information for an agent, and the costs of errors as likely to be high. For example, misjudging one's moral obligations could result in reputational damage or punishment, or even exclusion from the social systems. Fraser deems this condition is also unmet – at least based on available research so far.[214]

The error condition states that for an evolved cognitive mechanism to be reliable, it must be unlikely that the costs of errors are asymmetrical, as this tends to lead systematic bias via adaptive unreliability. By 'asymmetrical costs', Fraser means that the cost for false positives and false negatives differs greatly; that in some cases one of these kinds of errors can be more costly in terms of evolutionary fitness than the other. In such cases "selection can favour belief-formation mechanisms that overgenerate the less costly type of error so as to reduce the chance of a potentially disastrous error of the other type."[215] Assessing this in the case of the evolution of morality is very difficult, given the great variety of kinds of information and costs involved, and the lack of empirical evidence we have to evaluate them. At best, any attempt is likely be highly speculative. Fraser suggests the following example:

> Consider moral judgements about obligations to help others in danger or need. Helping in the mistaken belief that doing so is morally required incurs costs—e.g. time, energy, resources, or risk—but this kind of error could result in a net gain, if supererogatory helpfulness builds social capital. But even if such an error is on balance costly, failing to help in the mistaken belief that doing so was not required may be more costly still, if punishment and damage to reputation ensue. When it comes to helping others in danger or need, it may be better to save (and be saintly) than to stand by and be sorry. If this is right, then for this class of moral judgments, error costs are asymmetrical, and the error condition appears not to be met.[216]

[213] *Ibid.*
[214] *Ibid.,* p. 466.
[215] *Ibid.,* p. 462.
[216] *Ibid.,* p. 466.

It is hard to assess however whether this example is at all representative of error costs. Fraser's verdict is again that it is not clear if the error condition is met. Although he thinks there is some reasons to be doubtful, a safer conclusion is probably that there is a lack of information; there is no positive evidence to indicate the condition is met.[217]

The tracking condition is the familiar question of whether a cognitive mechanism's function is to track the relevant facts, or whether it was adaptive in some other way. Unlike the other three conditions, the tracking condition has been extensively discussed in the literature. As discussed in §3.3.2 and §3.4.2, there is not much agreement as to whether this condition can be met by realist accounts of morality. Fraser concludes, again tentatively, that the tracking condition is not met.[218]

Fraser's discussion addresses several empirical issues about the reliability of moral cognition that are under-discussed in the literature. His overall verdict is that if none of the conditions are conclusively met, "then it is reasonable to conclude that our moral faculty is actually unreliable."[219] However, the first three conditions he discusses are arguably mostly relevant in establishing the truth of the fourth: truth tracking. The differing environment, the informational costs, and the costs of informational errors in that environment are considerations that contribute to establishing that the function of the evolution of morality was not to track independent moral truths. Thus, Frasers considerations, while highlighting interesting empirical concerns and providing further reasons for taking arguments about evolutionary error theories seriously, do not significantly alter the landscape of the debate.

[217] *Ibid.,* p. 467.
[218] *Ibid.,* pp. 466-471.
[219] *Ibid.,* p. 472.

Chapter 4 Analysis of the implications of evolutionary arguments for ethics

In the previous chapter, I examined three separate attempts to argue from biological or evolutionary considerations to conclusions or positions in philosophical ethics. In this chapter, I review these arguments to see where they were successful, where they were not, and what we can learn from them. From looking at these attempts, I construct a framework to assist in assessing arguments of this empirical or evolutionary sort and to provide guidance for non-philosophers venturing into philosophical territory. As noted in chapter 1, the intended audience of this framework is non-philosophers working on interdisciplinary research involving philosophy, philosophers supplementing philosophy's methodology with the methods and tools of the sciences, or simply non-philosophers who discover their research appears to have philosophical conclusions.

4.1 Are Wilson, Joyce, and Street's arguments sound?

The success of arguments by E. O. Wilson, Richard Joyce, and Sharon Street that claim to extract philosophically significant results for ethics from the facts about the evolution of morality differs markedly. Wilson makes a number of claims, all of which are *prima facie* sensible in their approach. I argue however, that in some cases, mistakes are made and none of Wilson's four strands of 'biologicization' goes far enough in laying out the argument or in engaging with the actual problems of moral philosophy.

In contrast, Richard Joyce and Sharon Street take a much more thorough and distinctly philosophical approach to their arguments. While Joyce's evolutionary debunking argument may not by itself establish the strong form of scepticism or moral error theory that he perhaps aspired to originally, the argument effectively pushes the burden of proof on to the moral realist who claims moral naturalism is true, and challenges them to come up with a plausible account of moral naturalism that can

adequately meet the desiderata of the concept of morality. Sharon Street's Darwinian dilemma establishes that realist theories of value which include a claim of mind-independence are very likely to be false, and, to date, none of the published replies to Sharon Street's argument have successfully discredited its soundness.

As discussed in chapter 1, there is no simple rule of thumb, principle, or formula that can be used to assess arguments that attempt to derive philosophical significance from empirical considerations. However, it is unsatisfactory to state that all we can do is examine attempts on a case-by-case basis. To improve on this situation, in the sections that follow I examine the manner in which Wilson, Joyce, and Street have attempted to extract implications for moral philosophy from empirical information about the evolution of morality, explore the manner in which these arguments have been successful and unsuccessful, and explain what lessons we can learn from their attempts. The purpose of doing so is to use these lessons to come up with a framework that can be applied either when assessing such attempts or when attempting to make such arguments in the first place. It may not be feasible to build road through Philip Kitcher's swamp of empirical ethics in this thesis; however, it is possible to begin to lay some planks to avoid the most treacherous areas.

4.2 Lessons from E. O. Wilson

The four ways in which Wilson thought an understanding of Sociobiology would allow for the 'biologicizing' of ethics were:

1. The metaphysics of morality would be demystified and become apparent
2. An explanation of the problem of altruism would be provided
3. Our biology and evolutionary past might tell us what it is possible for us to do or be, and so constrains what we should do.
4. Whether something was natural would aid us in deciding its ethical status.

In the case of 1., an understanding of the on-going attempts at naturalising morality and the debates about the nature of moral reality and metaphysics is needed to contribute in a meaningful way to questions about the ontology and metaphysics of morality. Claims that some other discipline will be able to biologicize and thereby demystify ethics by removing it from the hands of moral philosophers should be viewed with scepticism if there is not considerably more that can be said by way of argument. Metaethics is a complex discipline in its own right and many metaethical problems are non-obvious, as is the understanding of them that has developed dialectically over long periods. To ignore this body of already existing understanding is effectively to start from scratch again. Indeed, this point applies to all of 1. to 4.: the problems identified by Wilson are all problems that have long been identified by moral philosophers (naturalising morality, explaining altruism, can-not implies ought-not, equating naturalness with goodness).

In the case of 2., Wilson conflates a problem of biology with a related, but not identical, problem in philosophy. The problem of explaining biological altruism via evolutionary mechanisms was resolved in the 1960's and 1970's and subsequent developments in game theory and genetics reinforced these discoveries. So, in the mid to late 70's when Wilson published his *Sociobiology* and *On Human Nature*, it was perhaps natural to think that these explanations may be novel to other areas of inquiry and, therefore, also solve problems in other disciplines. However, the problem of altruism in philosophy is importantly different to the biological problem with the same name. Wilson either does not acknowledge or does not recognize these differences. This means his argument is unlikely to be of interest to philosophy, as the fact that people can be altruistic (in the non-biological sense) was already long established. The fact that biological altruism has an evolutionary explanation does not contribute anything further to the philosophical concept.

The third way in which ethics may be 'biologicized' has promise; our biology and evolutionary past might tell us what it is possible for us to do or be, and so constrain what we should do. However, Wilson's claims of limits for our nature and behaviour is more an indication of the direction one might

make an argument about rather than a sufficiently elaborated argument itself. There is obviously a sense in which Wilson's claim is true: cannot implies ought not; if we are unable to carry out an action, this implies that it cannot be the case that we ought to perform the action. However, recognition of this fact is not new. And, Wilson fails to provide examples which would allow for a proper analysis. If the claim is that X is not possible for humans or societies due to some feature of their biological nature, then it is impossible to assess this argument without knowing what X is.

This type of reasoning is also very vulnerable to being used to push a particular moral or political agenda and is open to all the criticisms made against biological determinism. That is, that we need to be very careful in claiming anything as strong as a kind of genetic determinism when talking about human behaviour, as behaviours are the result of complex interactions between both biology and the environment. It is very hard to establish what is possible based on what currently is the case.

Finally, for point 4., and Wilson's argument that discovering whether something is natural or not could help determine its ethical status, there is no need to labour this point: some arguments are simply not sound. It is worth noting that it is well accepted in philosophical literature that naturalness is not an infallible guide to goodness, and if input had been sought from philosophers, there would no doubt either be a much more interesting argument that might show why this should not be dismissed immediately, or the argument would have been omitted.

4.3 Lessons from Richard Joyce's debunking argument

In contrast to Wilson's relatively unelaborated positions, Richard Joyce presents significantly better developed and more sophisticated arguments. An important feature of Joyce's arguments is that they have been revised numerous times over the years and been responsive to replies and criticisms of their content. Joyce's methodology is distinctly more philosophical and importantly takes as his starting point other positions in philosophy. He discusses potential and actual responses to his arguments, engages with the literature that already exists on the areas his arguments concern, and

references or extends similar ideas (for example John Mackie's arguments for error theory, and Michael Ruse's arguments about evolution). Joyce also makes clear the role of the empirical information used in his arguments. He argues that evolutionary theory and evidence shows that morality had a particular function, which was to produce moral beliefs. The moral beliefs themselves contributed social cohesion and facilitated group living. In Joyce's argument the empirical information is ultimately information that uncovers facts about the epistemology of our moral beliefs – that they originate in concepts that have a causal history that is unrelated to truth.

4.4 Lessons from Sharon Street's Darwinian dilemma

Sharon Street also provides a good model of philosophical methodology for empirical or evolutionary approaches to moral philosophy. Street spends considerable time precisely delineating and defining the targets of her evolutionary arguments. Street argues that scientific accounts of the evolution of morality provide novel considerations that count against versions of moral realism that endorse mind-independent evaluative truth. By targeting her argument at very specific and well-defined philosophical positions, she does not attempt to conclude too much, and her argument has fewer potential weaknesses. Street also does not make giant leaps of reasoning and does not make any unexplained jumps from the descriptive to the normative. Instead, she argues from descriptive premises about what entities were present in the evolutionary account of the development of morality, and shows there is no way to introduce mind-independent value into that evolutionary account in a scientifically defensible manner. While this is a much more limited kind of evaluative skepticism than many evolutionary error theories or moral skepticisms, it is a more defensible argument because of it. Additionally, while its conclusion by itself may not be a revolutionary conclusion or one with significant normative significance, it can be used in more general or ambitious sceptical arguments as a premise.

In the course of her argument against mind-independent evaluative truth, Street deploys a number of arguments against moral naturalism. While these arguments cover much of the same ground as previous arguments in the philosophical literature, their flavour is distinctly empirical and focused on showing that mind-independent evaluative truth cannot be naturalised because it is in conflict with naturalism more broadly construed: it requires an unscientific account to introduce the mind-independence. Thus, one form of argument that she succeeds in using is to apply the epistemic standards of empirical research to philosophical arguments.

4.5 Developing a framework for evaluating empirical arguments in ethics

Drawing from the above considerations from each of the discussions of Wilson, Joyce, and Street, the following ideas are offered as guidelines for assessing empirical and evolutionary approaches to morality.

Lesson / Guideline	Case study / Source
Removing some area of philosophy entirely from the context of philosophical discourse, with the hope of revolutionising it with some new insight from another field of inquiry, is unlikely to be a successful project without the provision of good reasons for that rehoming of the problem.	§3.1.1 'The metaphysics of morality' Wilson argues ethics will be revolutionised by treating it as a biological problem but does not describe how this will happen or provide any reasons for thinking it would be successful. Removed from the context of the existing discourse it is hard to see what form this revolution takes.
Examine the philosophical literature to ensure that the problem that empirical research purportedly solves is a genuine problem of philosophy, and is not a seemingly similar, but philosophically uninteresting or unrelated, problem.	§3.1.2 'The problem of altruism' Wilson argues that altruism is possible, but the kind of altruism he argues is possible is a biological conception of altruism rather than a philosophically interesting moral one.

Lesson / Guideline	Case study / Source
Examine the current state of philosophical debate on the topic in question, both within literature and via engagement with philosophers with expertise in the particular area. There is not much to be gained (for philosophy anyway) in re-solving problems which philosophy has already moved on from.	§3.1.2 'The problem of altruism' It is well accepted that some forms of altruism exist. If Wilson had engaged with the philosophical literature on the topic, he would have recognised that the problem he was proposing to solve was already settled.
Ascertain the relationship between what you are attempting to argue and well-known philosophical rules, for example, rules about deriving an 'ought from is' or about the relationship between naturalness and goodness. If one's argument appears to be an exception to a particular rule, then examine exactly how and why it is an exception and make this explicit. Arguments should not attempt to rehash old or well-accepted positions unless they have something new to contribute to the debate or have discovered a clear problem with the received view.	§3.1.4 'Naturalness and morality' Wilson appears to leap uncritically from what is natural to what we ought to do without awareness of the existing literature or that philosophers generally consider this argument fallacious.
Present a full argument, including review and response from philosophers: one way to ensure an unsuccessful attempt to integrate evolutionary or biological considerations into philosophy is to have no engagement with pre-existing philosophical dialogue.	§3.1.1 'The metaphysics of morality' Wilson suggests that taking a scientific approach and "removing ethics from the hands of philosophers" will clarify the metaphysics of morality. However, he does not provide any indication of how or what this would look like. The arguments advanced here would all have benefited from engagement with the existing philosophical dialogue and ensuring they were complete enough to say something philosophically significant.

Lesson / Guideline	Case study / Source
Arguments need to be complete, including establishing the premises are true and how they lead to the conclusion. Where the conclusion is a general one, it is helpful to provide specific examples that demonstrate the general point, rather than simply asserting the general conclusion without context.	§3.1.1 'Naturalness and morality' Wilson's argument that the naturalness of certain behaviours is relevant to their ethical status was incomplete, and lacked both clear premises, a clear logical form, or any specific examples of how the ethical status was established.
Ensure that discussion includes potential responses to likely objections based on existing philosophical theories and actual objections from other philosophers and researchers.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce integrates evolutionary considerations into the existing philosophical debate on moral error theory.
Provide the context of the argument by discussing similar arguments, whether solely philosophical or also attempting to incorporate empirical considerations. Make clear how the argument differs from previous arguments.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce compares his arguments with those of Mackie and other error theorists to show where the similarities lie and how his argument differs.
Where possible, make clear the role of empirical information or empirical considerations in the argument in question.	§3.3 'Sharon Street's Darwinian Dilemma' Street discusses how our evaluative attitudes are shaped and thoroughly saturated with evolutionary forces – they would have not been recognisable as the kinds of things they are without that evolutionary influence. She is explicit about how the empirical facts constrain the possibilities for realism and the impact these have for the metaethical status of moral realism.

Lesson / Guideline	Case study / Source
Consider seriously potential and actual responses made to arguments based on both empirical and philosophical grounds and revise the position accordingly if necessary.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce has engaged in ongoing dialogue over the years and has revised his conclusions from the strong forms of moral scepticism to a more limited conclusion that moves the burden of proof to the moral realist and challenges them to come up with a plausible account of naturalism.
Applying epistemological standards from other disciplines can shed light on stubborn problems if done carefully. However, to do this, analysis is required to determine whether the epistemic standards are appropriate to the argument and philosophical work is required to analyse the implications.	§3.1.1 'The metaphysics of morality' Wilson attempts to apply scientific methodology to establish the existence of certain philosophically interesting phenomena – in this case altruistic behaviour. However, his analysis of the phenomena is insufficient meaning the argument targets an apparently similar but philosophically less interesting sense of the concept altruism.
Often when empirical considerations are applied to moral philosophy, the goals of the arguments are sweeping and revolutionary. It is easy to overlook sound and potentially noteworthy philosophical conclusions that are revealed when arguing for these sweeping or revolutionary goals. These limited and more easily established and defended conclusions can often themselves be premises or assumptions in other philosophically interesting arguments or positions that are overlooked by focusing on the more revolutionary or dramatic conclusions.	§3.2 'Richard Joyce's evolutionary debunking argument' §3.3 'Sharon Street's Darwinian dilemma' These arguments initially tried to show that moral error theory or scepticism was true and that the entire practice of ethics needed a fundamental re-evaluation. However, while these conclusions may not have been conclusively established, more limited and potentially interesting considerations about certain aspects of moral naturalism, such as the mind-independence of moral truth or the contingent nature of morality, have been uncovered as part of those arguments.

Chapter 5 Moral psychology

5.1 Models of moral judgment

How people make moral judgments has been the subject of investigation in primarily two disciplines: philosophy and psychology. The approach of these two disciplines however differs significantly. Moral philosophy's interest in how moral judgments are made has often been an instrumental one: assumptions and claims about the process of making moral judgments have been made as part of wider meta-ethical accounts. Nevertheless, philosophical treatments of moral judgment often make many assumptions or claims about how we do in fact make moral judgments.

These philosophical accounts typically rely on *a priori* arguments or anecdotal examples and observations. Studies of moral judgment within psychology have tended to be more concerned with *a posteriori* methods, but often also included some *a priori* assumptions or starting points as in philosophy. Early research done on the moral development of children by Lawrence Kohlberg for example was empirical in nature; it examined people's actual responses to moral dilemmas to ascertain how judgments are made at different stages of a child's moral development. However, an unquestioned assumption in this work was that the process producing moral judgments was always kind of formal reasoning, and individuals were only judged as being competent at each stage if the justifications they gave showed reasoning of a pre-defined kind; a rational thought process based on prudential, conventional, and universalizing principles.[220] This historical separation between different methodologies is not however absolute and many researchers from both philosophy and psychology have begun to take a more interdisciplinary approach to moral psychology in the last few decades.[221] The focus of this chapter is on the empirical approach to moral judgments: what science can tell us about how we make moral judgments.

[220] John Doris, 'Moral psychology: Empirical approaches', sec. 1.
[221] *Ibid*.

The structure of this is as follows. Firstly, I present three simplified models or processes that are involved in most accounts of the production of moral judgments. These models I call the 'Rationalist model', the 'Emotivist model', and the 'Intuitionist model'.[222] The terms for them have been chosen so as to be descriptive of the dominant process in each model; they do not directly correspond to any ethical theories in philosophy with the same or similar names. These three models are used, either exclusively or in various combinations with different weights given to each component, in most accounts of how moral judgments are made. Following this, I consider some of the research that has been done to establish how these models should be combined to produce a realistic account of how moral judgments are actually made.

5.1.1 Rationalism

Rationalism in moral judgment, as I shall use the term, is where moral judgments are produced by a process of *reasoning*. When some morally relevant situation or event takes place, the details of the situation are used as inputs to a process of explicit reasoning involving conscious deliberation which subsequently produces a moral judgment (by 'conscious' it is meant that "the process is intentional, effortful, and controllable and that the reasoner is aware that it is going on"[223]). Generally the process of reasoning that occurs is said to be some kind of assessment of whether actions meet or violate a set of rules, or a kind of weighing up of the evident consequences and likely future consequences (perhaps as in utilitarian or consequentialist ethical theories), or even possibly the application of appropriate general moral principles regarding whether moral duties and obligations have been met by the participants in the eliciting situation (as in a kind of deontological model). Alternatively, it might also be taken to be a more traditional Kantian model where actions are examined to see whether they

[222] This is loosely based on Marc Hauser's terminology in *Moral minds: How nature designed our universal sense of right and wrong,* as they turn out to be fairly accurate representations of the elements of most models of moral judgment.
[223] Jonathan Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', p. 817.

are appropriate candidates that could serve as a universal law or maxim. This very simple model might look something like the following:

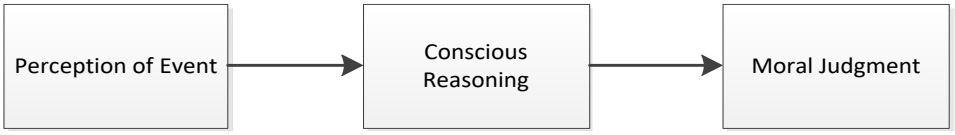


Figure 10: Rationalism in moral judgments

Early work in psychology on moral judgment was focused mostly on the process of conscious reasoning as the dominant process in arriving at moral judgments. Prominent in early research was the work of Jean Piaget and Lawrence Kohlberg. Their view of moral judgments was that they were "based on the ability to reason through the terrain of moral dilemmas, concluding with a judgment that is based on clearly defined principles."[224] As children developed morally, they acquired access to different relevant moral principles at each stage. The earliest stages involved principles to do with parental authority, punishment, and obedience, moving through to later stages which focused on things such as basic rights, the legal arrangements of a society, and ultimately adherence to universal ethical principles at the most morally developed stage. This was the default view of the development of moral judgment in psychology for a long time, however recent work has begun to focus on other elements thought to be involved: emotion and intuition.

5.1.2 Emotivism

The emotivist model is based on the idea that judgments of rightness and wrongness are arrived at primarily as reactions to emotions. The roots of this model can be seen in the work of David Hume who claimed that our "taste" (here meaning our perception of our emotions: our sentiments, likes, dislikes, desires, feelings and so on) "has a productive faculty, and gilding and staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation."[225]

[224] Marc Hauser, *Moral minds: How nature designed our universal sense of right and wrong*, p. 16.
[225] David Hume, *An enquiry concerning the principles of morals,* p. 88.

The basic idea here is that when we see some morally pertinent situation – a dog being beaten say, to use a Humean example – it arouses in us feelings of sympathy, and perhaps anger or disgust at the perpetrator of the beating, and these feelings give rise to Hume's 'new creation' – a judgment of moral wrongness of the situation. Thus, the emotivist picture of moral judgment is that we feel an emotion response to our perception of some event. We interpret the actions involved in the event as right or wrong depending on the emotional response and express this as a moral judgment. As Marc Hauser writes for example "in the same way that we automatically and unconsciously see red, hear music, smell perfume, and feel roughness, we perceive helping as right because it feels good and cheating as wrong because it feels bad."[226] A representation of the emotivist model in its simplest form is as follows:

Figure 11: Emotivism in moral judgment

5.1.3 Intuitivism

Intuitionism is the idea that our moral judgments are the result of special kinds of intuitions. Intuitions are differentiated from reasoning in that they are produced almost instantaneously, and do not involve any kind of explicit evaluating or conscious deliberation. Such judgments are generally characterized as appearing "suddenly and effortlessly in consciousness, without any awareness by the person of the mental processes that led to the outcome" and that the judgment process "does not advance in careful steps rather, it involves manoeuvres based seemingly on an implicit perception of the total problem"[227] Intuitions simply come into the mind unbidden as a response to some event that has been perceived. That moral judgments are the result of intuitions was a popular thought in early

[226] Marc Hauser, *Moral minds: How nature designed our universal sense of right and wrong*, p. 24.
[227] Jonathan Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', p. 818.

twentieth century philosophy, with theorists such as G.E. Moore, W. D. Ross, and H. A. Pritchard putting forward various kinds of ethical intuitionism.[228] The corresponding basic model for intuitionism is as follows:

Figure 12: Intuitivism in moral judgment

How intuition is cashed out here varies. In philosophical ethical intuitionism, intuitions were sometimes said to be a species of belief, the truth of which is somehow simply self-evident, or that they are a kind of *sui generis* insight that apprehends the moral features or properties of the world. A more recent take on intuitionism, in psychology in the work of Marc Hauser, characterizes intuition as the output of an evolved moral faculty, that produces judgments of rightness or wrongness in a manner that is analogous to how people judge whether a sentence is grammatical or not.

5.2 Composite models of moral judgment

All three processes of the §5.1 are easily recognizable as features of our moral lives. Firstly, moral judgments can clearly involve reasoning: moral philosophy is evidence that this can and does occur. One might wish to argue otherwise, however, if it is not admitted that at least *some* of the time moral philosophy does involve reasoning, then it will be hard to imagine *anything* that would satisfy the definition of a process involving 'reasoning'. Even if the other elements are present, it would be hard to deny that some reasoning and rational deliberation do occur in making at least some moral judgments. Secondly, emotion is a persistent feature of moral judgments. Morality is important and relevant to us in virtue of the fact that it involves things that we care about or feel strongly about:

[228] Seminal statements in: G. E. Moore, *Principia ethica*, H.A. Prichard, 'Does moral philosophy rest on a mistake?', W.D. Ross, *The right and the good.*

anger, gratitude, indignation, disgust, sympathy, contempt, shame, guilt, pride[229] – to name a few – are keenly felt when situations of a genuinely moral nature present themselves to us. Thirdly, something like intuition also appears to play a regular part in arriving at moral conclusions: judgments often seem to appear suddenly in consciousness, "without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion"[230]. These intuitions are not themselves reasoning, but they do not appear to be emotions either: they produce output that is belief-like, similar to that of reasoning, rather than simply an affective state. Any adequate account of moral judgments must be able to explain how these elements fit together in producing moral judgments. In the following sections I look at a number of theorists who have attempted to produce such an account, based on either their own research, or a collation of other studies done in psychology.

5.2.1 Greene et al.'s model

The first of the models that I look at comes from the work of Joshua Greene.[231] In Greene's study on moral judgments, participants were presented with a number of moral dilemmas which they had to make judgments about, while researchers administered fMRI[232] scans. The moral dilemmas presented to participants were split into two varieties: 'personal' and 'impersonal' moral dilemmas. An example of an impersonal dilemma is the well-known trolley problem:

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley

[229] List of emotions from Richard Joyce, *The evolution of morality*, p. 94.
[230] Jonathan Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', p. 818.
[231] See J. D. Greene, R. B. Sommerville, L.E. Nystrom, J. M. Darley, and J. D. Cohen, 'An fMRI investigation of emotional engagement in moral judgment'.
[232] fMRI or Functional magnetic resonance imaging, looks at changes in blood flow and blood oxygenation in the brain to determine regions of increased neural activity.

> onto an alternate set of tracks where it will kill one person instead of five. Ought you to turn the trolley in order to save five people at the expense of one?[233]

And an example of a personal dilemma used is a modified version of the above dilemma called the 'footbridge problem':

> As before, a trolley threatens to kill five people. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Ought you to save the five others by pushing this stranger to his death?[234]

The purpose of carrying out the fMRI scans while the participants made judgments about the dilemmas was to reveal which areas of the brain were active during the processing of personal and impersonal moral dilemmas and to evaluate the hypothesis that "the crucial difference between the trolley dilemma and the footbridge dilemma lies in the latter's tendency to engage people's emotions in a way that the former does not"[235]. The results showed that the areas of the brain usually associated with emotion were much more active during consideration of personal dilemmas compared to impersonal dilemmas. Additionally, the actions in the personal dilemma were more often judged to be less permissible than the impersonal dilemmas and when the events in the personal dilemmas were judged to be permissible, these judgments took significantly longer to make. Green hypothesized that the increased duration in such judgments was due to a "countervailing emotional response" that caused a kind of "emotional interference"[236] in such cases. Thus, the picture of moral judgments that we get from Greene's research is that judgments involving 'personal' situations are either influenced by, or produce a greater level of affective response, whereas impersonal moral dilemmas produce

[233] *Ibid*., p. 2105.
[234] *Ibid*.
[235] *Ibid*., p. 2106.
[236] *Ibid*.

judgments mostly through reasoning. While, Greene et al. do not provide a complete model of judgment themselves, a model based on their work would look something like the following:

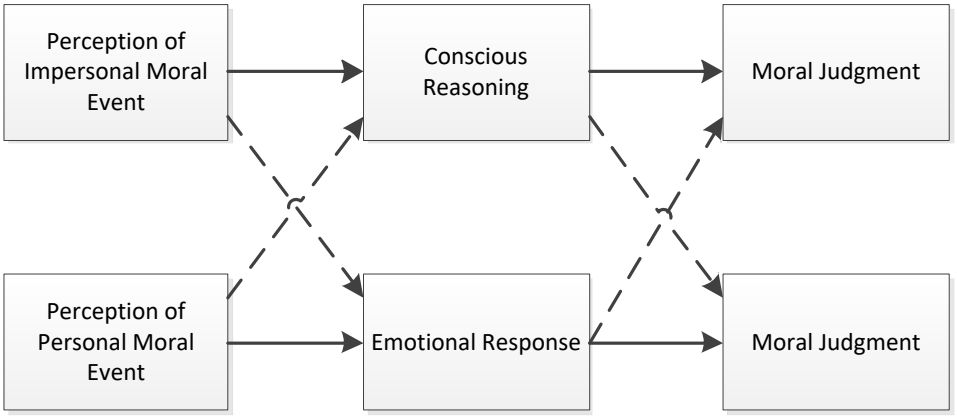


Figure 13: Greene et al.'s model of moral judgment

In this model impersonal moral situations do not trigger emotional responses of the same level as personal moral situations, and therefore judgments of the impersonal kind are arrived at primarily as a result of conscious reasoning. In personal moral dilemmas, emotional systems are triggered and play a major role in determining the outcome of the judgment. The links to conscious reasoning in personal moral judgments, and emotional response in impersonal moral judgments, are dotted to indicate the decreased role they play in such judgments. Thus, Greene's study provides some initial evidence for thinking that 'personal' and 'impersonal' moral dilemmas are processed differently, and that emotional and reasoning centers of the brain are involved to varying degrees in making moral judgments.

5.2.2 Shaun Nichols' sentimental rules model

In his book *Sentimental Rules: On the natural foundation of moral judgment,* Shaun Nichols presents a model he calls the 'sentimental rules account of moral judgment'. His account is limited to what he calls "core moral judgment" – judgments of the kind that are easily distinguished as moral, even by

children as young as 3 years, in typical moral/conventional task psychological studies.[237] For Nichols, the capacity for moral judgment is one that allows us "to recognize that harm-based violations [of moral rules] are very serious, authority independent, generalizable and that the actions are wrong because of welfare considerations."[238] It is important to note that while this may capture a significant part of what we refer to as 'moral judgments', it certainly does not exhaust the scope of what we use the term 'moral' to refer to.[239] To take one of Nichols' own examples, we generally think that there is something morally wrong with tax avoidance or fraud[240] (benefiting from other's taxes, while not contributing oneself even though one meets the qualifying criteria), but this kind of case does not trigger the directly harm-based type of judgment implicated in the moral/conventional distinction and thus is a moral judgment that is not covered by Nichols' model of how moral judgments are made.

Nichols' account of the process of moral judgment involves both an emotional response (an 'affective system') and a kind of reasoning about transgressions of an internalized set of rules – what he calls a 'normative theory'. By normative theory, Nichols means something very basic and is not intended in any inflated sense. Rather, even a basic set of rules prohibiting certain behaviours will count as a normative theory. Internally represented rules concerning table manners, for instance, will count as a normative theory."[241] The relation between the two mechanisms is not fully specified, but he argues that the affective response adds moral weight to the system of rules, as he puts it "the affective response infuses the harm-norms with a special non-conventional status."[242] He argues that an informative parallel is the disgusting/conventional distinction. Moral norms are a set of norms about harm transgressions that are backed by an affective system whereas conventional norms lack this

[237] As first exemplified in the research of Elliot Turiel, see Shaun Nichols, *Sentimental rules*, p. 5.
[238] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 7.
[239] Indeed, a recent article has highlighted the fact that all of the studies Nichols uses to support his theory employ a very narrow range of moral/conventional transgressions: those of the "sort that primary school children might commit in the schoolyard" – even when participants in the study were convicted adult psychopaths in prison. See Daniel Kelly, Stephen Stich, Kevin J. Haley, Serena J. Eng, and Daniel M. T. Fessler, 'Harm, Affect, and the Moral/Conventional Distinction'
[240] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, pp. 6-7.
[241] *Ibid*., p. 16.
[242] *Ibid*., p. 29.

emotive component. Likewise there is also a set of 'disgust norms', which have an affective element that go along with them that distinguish them from conventional norms. Breaking disgust norms is viewed as more serious, authority independent, and applicable to all contexts (generalizable) and the reason for this is that not only are people aware that there are rules against such actions, but that we have an affective response as well (i.e. transgressions just seem gross) and therefore appear to be more serious regardless of how well entrenched or explicitly important the rules about them are, authority independent because the affect does not disappear even if an authority excuses a transgression, and generalizable because the emotional response is produced reliably, regardless of the situation (something that is gross, is gross regardless of the context). So, in the same way as disgust norms are backed by emotive responses, so are moral norms, and this leads us to treat them as distinctive and as having the features described above.

Thus Nichols proposes that moral judgments are made by assessing whether particular actions or situations are transgressions of some set of shared rules, and additionally, they get their importance and special status from the affective reactions that accompany them. Nichols thinks that moral judgments are not made by any special kind of 'moral sense', but are instead simply a particular kind of judgment about norms that have developed through a process of cultural evolution. Traditionally a number of accounts of moral judgment have included a unique "moral sense as the source of distinctive feelings of approval and disapproval which are triggered by the perception of virtue or vice".[243] Nichols however argues that no such unique moral sense is necessary or present: on his account all that is needed is the accompanying affective reaction which "plays a crucial role in leading people to treat harmful transgressions as wrong in a distinctive way."[244] Basic emotions from other contexts supply the sentiment to moral judgments and "No further moral feeling is invoked as a necessary part of core moral judgment."[245] Nichols' model looks like the following:

[243] *Ibid*., p. 62.
[244] *Ibid*., p. 63.
[245] *Ibid*.

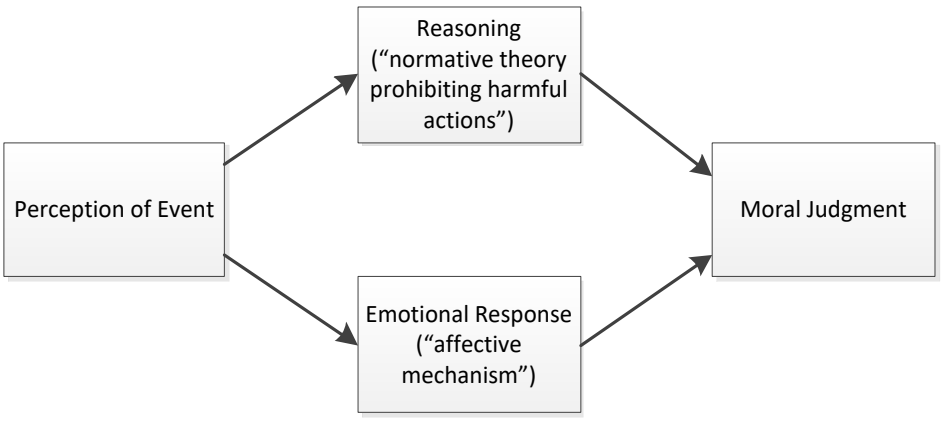


Figure 14: Shaun Nichols' sentimental rules model of moral judgment

Nichols provides support for his model by using results from studies on the moral/conventional task and research into the disgust/conventional distinction. In these studies, participants are asked to judge whether a transgression is of a moral or conventional nature: the studies explicitly ask whether a transgression has taken place. The first question in standard versions of the task checks for the permissibility of an action – and thus whether it breaks some set of rules or norms. So, typical moral judgments involve assessing whether a norm has been broken. Further, there is, in moral/conventional tasks, a probe that is often used to distinguish between transgressions and non-transgressions which involves asking the subject whether punishment is appropriate for certain events. Nichols gives the examples of a natural disaster and a child falling over and skinning their knee – both of these are bad, but neither make sense to punish for – they do not involve transgressions.[246] Thus in Nichols' model, judgments that something is *morally* bad or good involve an assessment of whether a transgression has taken place.

The evidence for the second half of his model, that moral harm norms are backed by an affective mechanism that gives them their special moral features, is drawn from two sources. Firstly, studies done on psychopaths[247] show that while psychopaths possess the capability to recognize that moral transgressions involve breaking rules or expectations, they lack the ability to distinguish moral

[246] *Ibid*., p. 15.
[247] Such as R. J. R. Blair, 'A cognitive developmental approach to morality: investigating the psychopath'.

transgressions involving harmful actions from conventional ones.[248] Additionally psychopaths' judgments lack any emotional content: they do not have any affective response to the suffering of others involved in cases of moral transgressions involving harm.[249] Secondly, Nichols uses studies on disgust which show that norms prohibiting disgusting actions are treated as more serious, authority independent, and universal (features which moral norms also have compared to conventional rules) when they are accompanied with an affective response to the disgusting action. As Nichols writes "If we find that other affect-backed norms [such as disgust norms] are also distinguished from conventional norms along the dimensions of permissibility, seriousness, authority contingency, and justification type, then this will provide an independent source of evidence for the Sentimental Rules account."[250] And this is precisely what two studies on disgust violations undertaken by Nichols do show.[251]

5.2.3 Jonathan Haidt's social intuitionist model

Jonathan Haidt has proposed what he calls the 'Social intuitionist' model of moral judgment.[252] The thoughts motivating this model are that moral *reasoning* is not the main cause of moral judgments, but is instead usually a *post hoc* construction, generated after a moral judgment has been reached, and that moral judgment regularly has a social element. Haidt's somewhat complicated looking model is as follows[253]:

[248] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 19.
[249] *Ibid*., p. 19.
[250] *Ibid*., p. 21.
[251] See Shaun Nichols, 'Norms with feeling: Towards a psychological account of moral judgment'.
[252] Jonathan Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment'.
[253] *Ibid*., p. 815.

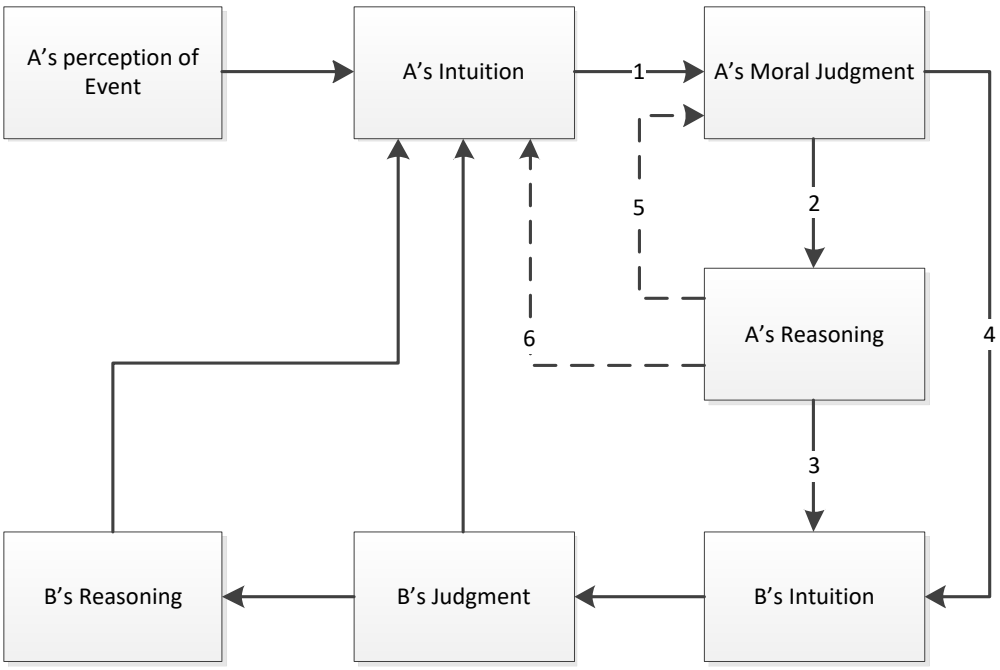


Figure 15: Jonathan Haidt's model of moral judgment

In Haidt's model, moral judgment is most usually the result of an intuition (link 1 in the model, called the "intuitive judgment link"). Reasoning according to the model is engaged in after a judgment is made, to search for arguments that will support the conclusion (link 2, the "post hoc reasoning link"). The third link (the "reasoned persuasion link") consists of moral reasoning that is produced and sent forth verbally to others to justify one's moral judgments. While this link is sometimes efficacious, discussions about moral judgments are "notorious for the rarity with which persuasion takes place."[254] The fourth link in the model, is the "social persuasion link." This link comes from the recognition that the mere fact that other people – friends, family, allies, and acquaintances – have made a moral judgment "exerts a direct influence on others, even if no reasoned persuasion is used."[255] These four links in the model constitute the core of Haidt's model: moral intuitions, post hoc reasoning, and influences from our social interaction about moral judgments. The fifth and sixth links in the model are hypothesized to occur less frequently and be less influential in causing judgments. This is indicated in the model by the use of dashed lines. The fifth link, the "reasoned judgment link" is where "people

[254] *Ibid*., p. 819.
[255] *Ibid*.

reason their way to a judgment by sheer force of logic."[256] This link recognizes that in rare cases people do engage in genuine, non-*post hoc* reasoning about judgments they have already made, and this results in a revision or change of judgment. This most often happens when the initial intuition is weak or inconclusive. The "private reflection link" is the sixth link in the model. This is where in the course of reflecting on the situation or issue, a person may initiate an intuition that differs from their original intuitions that caused their judgment.

Haidt offers four reasons for doubting the causal efficacy of reasoning in moral judgments and preferring his social intuitionist account. The first is that it is widely accepted in psychology that other kinds of judgment often involve two kinds of processing systems (a rapid intuitive or automatic system and a slower reasoning based one), usually called *dual process* models, and he thinks there is good reason to think moral judgment works in a similar way.[257] Normally in such models, the affective or intuitive system "has primacy in every sense: it came first in phylogeny, it emerges first in ontogeny, it is triggered more quickly in real-time judgments, and it is more powerful and irrevocable when the two systems yield conflicting judgments"[258] Haidt thinks that evidence from relevant studies show judgments to be best described as "a set of automatic processes [rather] than as a process of deliberation and reflection"[259], that people's "impressions that they form from observing a 'thin slice' of behaviour (as little as 5s[econds]) are almost identical to the impressions they form from much longer and more leisurely observation"[260], and that people "categorize other people instantly and automatically, applying stereotypes that often include morally evaluated traits."[261] All of the findings illustrate an intuitive process whereby "the perception of a person or an event leads instantly and automatically to a moral judgment without any conscious reflection or reasoning."[262]

[256] *Ibid*., p. 819.
[257] *Ibid*.
[258] *Ibid*.
[259] *Ibid*., p. 820.
[260] *Ibid*.
[261] *Ibid*.
[262] *Ibid*.

Haidt's second reason for questioning the rationalist picture of moral judgment is what he calls the "motivated reasoning problem" – that the moral reasoning process seems more like a lawyer defending a client than a judge or scientist seeking the truth.[263] That is, the reasoning process' purpose seems to be to formulate arguments that support one's intuitive conclusions rather than having a role in determining the outcome of the moral judgment itself. In support of this claim, Haidt cites a number of studies which show that reasoning is often biased by motives concerning 'impression management' and 'smooth interaction' with others (what Haidt calls "relatedness motives"), and defensive mechanisms that are triggered by threats of incoherence or invalidity of one's views (so called "coherence motives"). These mechanisms make people "act like lawyers" and defend their claims through any reasoning they can. Although Haidt thinks that sometimes moral reasoning may be efficacious at working to produce moral judgments, this occurs only in very limited circumstances. In most real situations, such as when people are arguing about a morally relevant situation, relatedness and coherence motives will underlie much of the reasoning: "under these more realistic circumstances, moral reasoning is not left free to search for truth but is likely to be hired out like a lawyer by various motives, employed only to seek confirmation of preordained conclusions."[264]

The third reason is that the reasoning process that constructs 'rational' justifications for intuitive judgments often produces reasoning that is inadequate in terms of explaining the judgment. For example, in a study done by Haidt,[265] a fictional case of consensual incest between a brother and sister on holiday is described to participants. Most people immediately judge the actions wrong, but struggle for sufficient justifications for their judgment when questioned. Eventually participants end up claiming that the brother and sister will end up harming themselves emotionally despite the fact that the story told to them intentionally rules out any possible harm to the participants. Ultimately many people end up saying something like "I know it's wrong, but I just can't come up with a reason why."[266]

[263] *Ibid*., p. 820.
[264] *Ibid*., p. 822.
[265] J. Haidt, F. Bjorklund, S. Murphy, 'Moral dumbfounding: When intuition finds no reasoning'
[266] *Ibid*., p. 11.

Clearly this reasoning is not sufficient for rationally producing the judgment and it does not appear that the judgment could have been based upon reasoning at all. Haidt also thinks that this feature of the process can explain the "bitterness, futility, and self-righteousness"[267] of a lot of moral discourse or argumentation. Because the reasoning is usually not part of the making of judgments, rational arguments often fail to be persuasive, regardless of their soundness. As Haidt explains:

> In a debate about abortion, politics, consensual incest, or what my friend did to your friend, both sides believe that their positions are based on reasoning about the facts and issues involved… Both sides present what they take to be excellent arguments in support of their positions. Both sides expect the other side to be responsive to such reasons… When the other side fails to be affected by such good reasons, each side concludes that the other side must be closed minded or insincere. In this way the culture wars over issues such as homosexuality and abortion can generate morally motivated players on both sides who believe that their opponents are not morally motivated.

Haidt's fourth reason for thinking intuition is the cause of judgments and reasoning is a consequence, is that moral action co-varies with moral emotion more than with moral reasoning. While there is some evidence that moral reasoning is correlated with moral action, emotional and self-regulatory capacities seem to be much better determinants of negative morality – refraining from behaviour generally judged as immoral.[268] Additionally, positive morality – the active helping of others – is most likely precipitated by emotional reactions. Empathy aroused by the perception of suffering evokes altruistic motivations. Haidt sums up his review of the empirical support for this claim as follows: "people are often motivated to help others and…the mechanisms involved in this helping are primarily affective, including empathy as well as reflexive distress, sadness, guilt, and shame."[269]

[267] *Ibid*., p. 823.
[268] See *Ibid*., pp. 823-824.
[269] *Ibid*., p. 825.

5.2.4 Marc Hauser's Rawlsian model

Marc Hauser, in his book *Moral Minds*, presents a model of moral judgment which draws inspiration from John Rawls' *A Theory of Justice* where he compares our capacity to make moral judgments to our capacity to make judgments of the grammaticality of sentences. Rawls' insight was that our process of making judgments of right or wrong involves a kind of intuitive analysis, just as our process of judging whether a sentence is grammatical does. We make rapid, unconscious, and automatic analyses of the features of sentences, and produce near instant judgments of whether they are grammatically correct or not. Hauser thinks we analogously make rapid, automatic analyses of the moral features of situations and it is this analysis which produces moral judgments rather than emotional responses or reasoning. Once an individual perceives an action or event, the moral faculty provides a rapid analysis of the intentions and motivation underlying the action, its causes, and its intended and foreseen consequences. This non-conscious analysis then produces an intuition: a judgment that some action is permissible, obligatory, or forbidden.[270]

Building on the linguistic analogy, Hauser proposes that just as there is a kind of innate 'universal grammar' in linguistics, there is a kind of innate moral faculty that he calls a 'grammar of action'.[271] The linguistic universal grammar is the innate capacity or 'toolkit' that facilitates the learning of our native languages. Hauser suggests that humans are equipped with a parallel innate moral capacity: a universal 'moral grammar' – a toolkit for building our specific culture's morality. As Hauser writes "in the same way that grammaticality judgments emerge from a universal grammar of principles and parameters… ethicality judgments would emerge from a universal moral grammar, replete with shared principles and culturally switchable parameters."[272] Once we have acquired the moral norms of our culture, a process which, according to him is "more like growing a limb than sitting in Sunday school and learning about vices and virtues" we are able to intuitively judge whether actions are

[270] Marc Hauser, *Moral minds: How nature designed our universal sense of right and wrong*, p. 46.
[271] *Ibid*., p. 43
[272] *Ibid*.

"permissible, obligatory, or forbidden, without conscious reasoning and without explicit access to the underlying principles."[273]

One of the reasons Hauser thinks that there must be an intuitive process of analysis that is carried out before any emotion is felt is that something must tell us *which* emotion is appropriate to the situation. Hauser states this problem as follows: "Neither we nor any other feeling creature can just *have* an emotion. Something in the brain must recognize – quickly or slowly – that this is an emotion worthy situation."[274] In Hauser's model, the first step after the perception of the event that will result in a moral judgment is his intuitive analysis of the moral features of the situation. Once this analysis of the intentions and motivation underlying the action, its causes, and its intended and foreseen consequences has taken place, a moral judgment is arrived at. This judgment then gives rise to emotional responses and conscious reasoning about the situation. Hauser's model looks as follows:

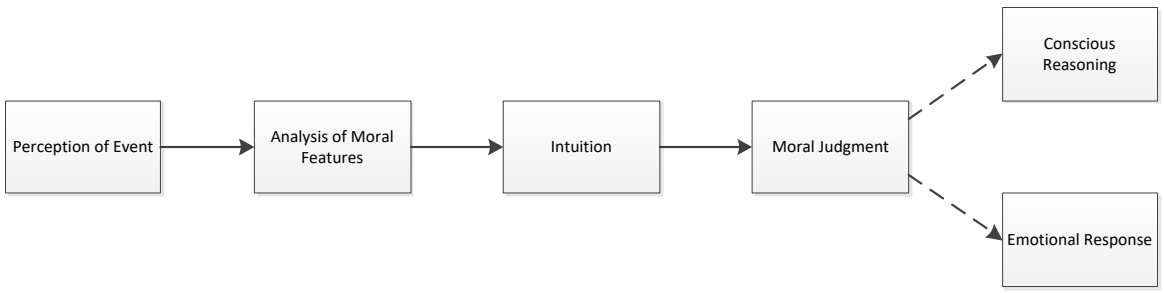


Figure 16: Marc Hauser's moral grammar model of moral judgment

In support of the claim that conscious reasoning typically follows *after* moral judgments have been made, Hauser cites similar research and evidence as Haidt does. His evidence for claiming that emotional responses follow from, rather than are causally involved in determining, moral judgments is somewhat limited.[275] It is worth noting that his argument that before any emotional response takes place there must be an analysis of what kind of affective response is appropriate, does not show that the affective response must take place after the moral judgment is made, only that it must be

[273] *Ibid*., pp. xvii-xviii
[274] *Ibid*., p. 8.
[275] See J. Nado, D. Kelly, and S. Stich, 'Moral Judgment', pp. 10-12.

preceded by *something*, and thus both the analysis and emotion could potentially come before the judgment.

5.3 Current models of moral judgment

The prior sections introduced four different empirically based psychological models of moral judgment. Most of the models that come from recent work in moral psychology focus on intuition or intuitive analysis of some kind as the primary way of arriving at moral judgments, although it is often also recognized that it is possible to arrive at moral judgments in more than one way (for example in Haidt's and Greene's models). Greene's model suggests that there is variation in the systems or sets of processes that produce moral judgments, depending on the eliciting situation and its relation to the individual making the moral judgment. Greene's study also indicates that more research on how different kinds of judgments are made is needed, and that it is unlikely that all judgments are made in the same way. Shaun Nichols claims in his model, based on moral judgments exemplified by those used in moral/conventional task research, that there are two kinds of processes necessary for making judgments. The two processes he suggests are an analysis (which could be a form of conscious reasoning or some kind of subconscious analysis), about whether a transgression has taken place, and an affective response that infuses the judgment of the transgression with its importance, authority independence, and universalizability. Jonathan Haidt's social intuitionist model of moral judgments is more complex and contains influences from the moral judgments of other individuals as well. However, in Haidt's model the dominant process is what he calls the "intuitive judgment link" – a kind of intuition. Judgments, according to Haidt, can be revised through reasoning and private reflection, reasoned persuasion by others, and also brute social biases such as what he calls "coherence" and "relatedness" motives. Marc Hauser's model is yet another model that explicitly favours intuitions. These, according to Hauser, take the form of a rapid, unconscious analysis of the situation, which results in a judgment that an act or event is "morally permissible, obligatory, or forbidden". Explicit

reasoning and emotional responses follow after the moral judgment in Hauser's model, although there is some parallel with Haidt's in that he thinks the function of these is both to tip the weight of some moral dilemmas one way or the other, and to provide reasoning to others, often with the aim of changing their judgment.

5.4 How are moral judgments made?

While this research shows much promise and is often insightful, it is still relatively exploratory and for the most part the empirical evidence at present does not allow for a complete picture of the processes involved in moral judgment or any definite conclusions about how people typically make moral judgments in specific situations. However, such work is useful for showing where more research is required and the paths of enquiry that might be fruitful. The models that have been presented provide relatively crude characterizations of the process of moral judgment, and some of them disagree upon which processes should be included in the model or where the different processes should be located. Further, they tend not to take into account variations in the eliciting situation or any other variables dependent upon the nature of the moral dilemma that is being judged. Nevertheless, the research to date has shown that answering the descriptive question of "how are moral judgments actually made?" is tractable using the empirical methods of psychology, and gives us sufficient reason to think that more accurate models will be produced that provide a more robust and complete account of moral judgment.

Chapter 6 Moral psychology and ethics

A number of moral philosophers have taken an interest in the empirical findings of moral psychology, and claim that they have significant implications for ethics. While there is still an often-mentioned rule of thumb in ethics that one cannot directly derive normative prescriptions from descriptive information, this need not rule out moral psychology having implications for the less obviously normative field of *meta-ethics* – the study of the meaning, nature, and foundations of ethics. And further, there are some philosophers who do argue that there are significant normative conclusions that can be drawn from the findings of moral psychology. For example, Shaun Nichols, an enthusiastic proponent of applying psychological data to ethics writes, "it turns out that there is a great deal of extant empirical work that is philosophically instructive"[276] and that his psychological account of moral judgment which I discussed in the previous chapter, "has broad ramifications for philosophical ethics."[277] Even more sweepingly he states that "many of the deepest issues concerning the nature of morality would be illuminated if we had an adequate account of the nature of moral judgment."[278] Similarly John Doris and Stephen Stich write that "consideration of work in the biological, behavioural, and social sciences promises substantive philosophical contributions" and that "philosophical ethics can, and indeed must, interface with the human sciences."[279]

Moral psychologists themselves are usually reserved in their claims about the direct consequences of their research for moral philosophy. They do however comment on how a *lack* of understanding of moral psychology has somehow resulted in the failure of moral theorising to have an impact on people's actions, and that if we want to make decisions and policies that are successful, we should pay attention to what moral psychology has to say. For example, Marc Hauser writes:

> The dominant moral-reasoning view has generated failed policies in law, politics, business, and education. I believe that a primary reason for this situation is our

[276] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. vii.
[277] *Ibid*., p. viii.
[278] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 4.
[279] Doris and Stich, 'As a matter of fact: Empirical perspectives on ethics', p. 146.

> ignorance about the nature of our moral instincts and about the ways they work and interface with an ever-changing social landscape. It is time to remedy this situation."[280]

Similarly, Jonathon Haidt writes "A correct understanding of the intuitive basis of moral judgment may therefore be useful in helping decision makers avoid mistakes and in helping educators design programs (and environments) to improve the quality of moral judgment and behavior."[281] In addition to direct claims of significance from philosophers and psychologists there has been pressure in the last century within philosophy for ethical theories to be 'naturalistically acceptable': that they should cohere with, or 'fit' with, what science tells us about the natural world. Therefore, moral psychology has also been perceived to be increasingly important as it tells us, in part, what science has to say about the moral domain. An ethical theory that commends actions, obligations, commitments, and ways of living that are at odds with how people can realistically be expected to act or to be, according to science, is one that is increasingly at a serious competitive disadvantage. Since science tells us about the way people are (or importantly, if it can tell us about how people *can* be), then it can aid us in determining which ethical theories are viable and which are not.

While moral philosophy and moral psychology are both 'about' morality, their focus is of course, importantly different. While this may appear obvious, it is worth emphasizing, as where there is excitement about the implications of moral psychology there is often little discussion of what makes the methodological approaches and aims of each discipline dissimilar and these important facts sometimes get lost or forgotten in the ensuing discussion. Thus, I begin with a discussion of the ways in which moral philosophy and moral psychology differ with respect to their aims and methodology, and the impact these differences have on how moral psychology can contribute to ethics. There are however limits to the usefulness of discussing in the abstract how important the implications of moral psychology for ethics are, and thus following these general remarks I examine two different attempts

[280] Marc Hauser, *Moral minds: How nature designed our universal sense of right and wrong*, p. 2.
[281] Jonathan Haidt, 'The emotional dog and its rational tail: A social intiuitionist approach to moral judgement', p. 815.

to draw philosophical conclusions from work done in moral psychology. The first focuses on the work of Shaun Nichols who argues that the various findings from moral psychology on how moral judgments are made pose a challenge to various kinds of 'moral rationalism'. The second looks at an argument advanced by John Doris and Stephen Stich (among others) who argue that virtue ethics is committed to a picture of moral psychology that is at odds with the picture portrayed by psychology, and because of this virtue ethics is in serious trouble.

6.1 Methodological differences

Psychology is the scientific investigation of the human mind and human behaviour. Moral psychology is therefore the scientific study of the mind and behaviour of humans in moral contexts. Moral philosophy is less easy to define concisely for there is no charter or well-marked limits to what is an acceptable question for philosophical discourse other than the willingness to be taken seriously and discussed by philosophers. Nevertheless, moral philosophy has traditionally had a distinctive focus which differentiates its interests from that of moral psychology: it is concerned with what we *ought* to do; how we *ought* to live, how we *should* act, what policies and reforms we *should* support, what would be just or right or good, and so on. In short, the foremost difference between moral philosophy and moral psychology is that moral philosophy is *normative* whereas moral psychology is *descriptive*. Of course, moral philosophy is not limited just to discussing how we ought to act in any particular situation: there are many subtle distinctions involved in evaluating which judgments should be made, whether such judgments are true, or correct, or justified, or 'appropriate' in some other sense. These subsequent questions are questions about the *nature* of moral debate – questions that are usually considered to fall under the subject matter of metaethics.

Metaethics is *prima facie* non-normative or at least normative to a much lesser degree, and it is often thought that it is within metaethics that moral psychology is likely to be most relevant. However, it can be somewhat misleading to define metaethics as simply 'the study of the nature of morality', as

this makes it sound as though it is the kind of thing science can directly investigate – that metaethics might be itself a kind of descriptive domain. But the questions metaethics traditionally deals with are not obviously amenable to being dealt with in an empirical manner. When philosophers inquire about the nature of morality, they are asking questions such as the following:

- When we make moral judgments are we stating facts? If so, what kind of facts are they? Or, are we instead expressing our attitudes or desires?
- Can moral judgments be true or false? If not, can they be more or less 'correct' or 'appropriate'? How are they justified? *Are* they ever justified?
- How do we come to *know* that they are true or correct or justified? Can we have moral knowledge? What kind of epistemology is appropriate in the moral domain?
- Is there such a thing as objectivity in ethics? Or is ethics fundamentally subjective or relative to some agent or group of agents? Can we say what the objective/subjective distinction amounts to in ethics?
- Is there such a thing as moral progress? What does it consist in, and how do we know it has been made?

None of these questions (or other questions that fall under the umbrella of metaethics) appears to be easily addressed empirically. We should be cautious however of therefore assuming or jumping to the conclusion that empirical information or research is irrelevant to these questions.

James Rachels presents an analogy which attempts to show that biological and psychological information about morality is unlikely to be important, and on the basis of that analogy jumps to such a conclusion about the irrelevance of empirical information. His analogy is between mathematics and the study of the psychology of mathematics – that is, the study of people's mental processes involved in mathematical problem solving – and the moral philosophy and the study of the psychology of morality.[282] What would the study of the psychological processes involved in mathematical decision

[282] James Rachels*, Created from animals: The moral implications of Darwinism,* p. 78.

making contribute to solving mathematical problems? The answer is that mathematics and the psychology of mathematics tell us two very different kinds of things: one gives us solutions, the other tells us about the human mental processes and behaviour involved. Only by actually doing the problem-solving task can we hope to arrive at actual answers to mathematical problems.

Rachels points out that while information on *how* we solve mathematical problems will no doubt be of great interest to mathematicians, it is much less likely to help them solve any particular problem. For knowing which parts of the brain are involved, or what environmental or biological factors influence people's mathematical capacities, or just about any conceivable psychological information, does not contribute to the actual solution of any given mathematical problem. Similarly in the case of moral philosophy, while psychology's discoveries about moral judgments will undoubtedly be of great interest to moral philosophers, it is not clear how they could actually contribute to the problems of moral philosophy: deciding what to do and addressing all the philosophical questions the notion of "what to do" raises.

But this seems to be too fast. Simply supposing that because it is intuitively plausible to Rachels (or indeed any or all philosophers) that any descriptive information about the psychology of mathematics is not going to be helpful in solving actual mathematical problems does not mean we can therefore conclude that it is in fact of no use for mathematics. Further, Rachels' analogy is comparing the empirical understanding of human mathematical ability and the activity itself of mathematics with the empirical understanding of morality and the activity of moral philosophy (or normative ethics). A more relevant analogy for moral psychology and metaethics would be to compare the empirical study of the psychology of maths and its implications for the philosophy of mathematics. It seems likely that the philosophy of mathematics plays a similar role for mathematics as meta-ethics does for moral philosophy – if this is the case, then it would seem very hasty to draw the conclusion that Rachels has. If the philosophy of mathematics is the study of the assumptions, foundations, and implications of mathematics then discoveries in the philosophy of mathematics may well have implications for

mathematics (which few philosophers of mathematics would deny) and there needs to be more careful examination by those in a position to do so to conclude anything of the kind Rachels does.

There is at least one non-trivial[283] exception to this general rule about the relevance of psychological data to normative conclusions. This exception derives from the fact that it cannot be the case that one ought to do something if doing that thing is impossible – the principle of "ought implies can". Thus, it is possible from some descriptive data about the limitations of what is possible to do, that we can conclude that it is *not* the case that we ought to do something. If psychology discovers that some action, behaviour, or mental process is impossible – either for all people or some subset of people – then the result can be that we change our normative judgments about such cases. For example, if psychology shows that people who suffer from some disease or natural impairment are thereby rendered incapable of understanding that how they act is morally wrong, then we do not think that the deserts such actions merit are the same as the for those who act in a similar manner but *do* understand the nature of their actions. While this may appear to be a fairly limited exception – perhaps relevant only to cases involving a legal defence of pleading insanity or those without the cognitive capacities for understanding morality and so on – it can also be used in evaluating ethical theories as a whole. If it were true that some moral theory asks us to act in some way that is impossible or unrealistic to expect, then it casts doubt on the legitimacy of such a moral theory.

6.1.1 Terminological differences between disciplines

Before looking at some attempts to derive moral implications from psychological data, it is worth noting that there are differences in the way in which the same or similar terms are used between

[283] It is of course possible to derive trivial ethical propositions from descriptive propositions via various logical tricks. For example: "Tea-drinking is common in England; therefore either tea-drinking is common in England or all New Zealanders ought to be shot" (from A. N. Prior, 'The autonomy of ethics', p. 201.) Such statements are what Prior calls *contingently vacuous*: given the particular premises used to derive the conclusion, we could replace the moral predicate with any grammatically correct proposition and the truth value of the whole would remain unchanged. Thus, the ethical content in cases such as this are not really derived from the premises, and certainly this kind of logical derivation is no use to us in deciding what we ought morally to do.

disciplines, and the range of confusions this creates in attempting to draw philosophical conclusions from moral psychology. Moral psychologists often use the terms 'rationalism', 'emotivism', and 'intuitionism' to refer to the theses that moral judgments are the 'product' of the respective faculties. For example, Jonathan Haidt writes:

> Moral psychology has long been dominated by rationalist models of moral judgment…Rationalist approaches in moral psychology…say that moral knowledge and moral judgment are reached primarily by a process of reasoning and reflection.[284]

Similarly, the terms 'emotivism' and 'intuitionism' are used in psychology for the ideas that moral judgments are 'primarily the result of' or 'causally flow from' emotional or intuitive processes. Moral psychologists also sometimes term the psychological version of rationalism a 'Kantian' approach, and psychological emotivism a 'Humean' approach.[285] Generally, the use of all these terms does *not* map cleanly on to the way the same terms are used in philosophy. In moral philosophy, 'emotivism' is generally taken to be the view that moral judgments do not express beliefs, but instead function to express our feelings or attitudes when we make moral judgments. 'Intuitionism' is used in a number of ways, but is perhaps most often linked with the (now unpopular) epistemological and metaphysical view, that moral goodness and badness is apprehended by some form of intuition and is in some sense metaphysically *sui generis*. Similarly, 'rationalism' in moral philosophy is used in a large number of ways that differ from the sense used in psychology. The range of meanings attributed to moral rationalism in philosophy are too varied and complicated to summarise simply, but it suffices here to note that it does include the psychological idea that moral judgments are the result of a rational process, but also goes well beyond this to deal with issues such as justification (why we should act

[284] Jonathan Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment' p. 814.

[285] For example, Marc Hauser, terms his three possible models of the process of moral judgment 'Kantian' (judgments are made rationally), 'Humean' (judgments are the result of emotive responses), and 'Rawlsian' (judgments are the result of a moral faculty that judges intuitively whether an action is right or wrong based on what he terms a "grammar of action"), see *Moral minds: How nature designed our universal sense of right and wrong*, pp. 12-55.

morally) and what the concept of morality is about (that our concept includes the idea that moral requirements are requirements of reason or rationality).

Of course, the fact that different disciplines use the same terms in differing ways does not mean that work from different disciplines cannot be integrated. As long as the different meanings of terms is kept in mind and where there is overlap it is made clear which sense is intended this ought to pose little difficulty. As discussed in the following sections on the work of Shaun Nichols, this is not as easy a task as it may appear. Nichols' attempts run in to difficulty as many of his argument miss their philosophical targets due to their use of conceptions of terms such as 'rationalism' and 'moral judgment' that differ to those used in the relevant philosophical debates.

6.2 Shaun Nichols and moral rationalism

In *Sentimental Rules,* Shaun Nichols presents a psychological account of the elements involved in moral judgment (as outlined in Chapter 5), and argues that this model, and the psychological data that supports it, has important implications for ethics. Nichols examines how well his empirical model fits with what he calls 'moral rationalism' – the idea that "morality is grounded in reason or rationality rather than the emotions or cultural idiosyncrasies."[286] 'Moral rationalism' without any kind of qualification is somewhat ambiguous as this term is used in a number of ways in metaethics – it is applied to a range of ideas, some of which refer to particular traditions in moral philosophy, the ideas of various groups of philosophers, or sometimes it is applied generally to just about any moral theorising which connects reasoning or rationality in some way with morality. Nichols describes the

[286] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment,* p. 66.

basic idea of moral rationalism simply as the idea that morality is "founded on reason or rationality"[287], "derives from reason"[288], is "grounded in reason or rationality"[289], or that it "is a product of reason"[290].

Nichols suggests two different theses that he takes the philosophical moral rationalist to be advancing by these kinds of claims. The first he terms 'Empirical rationalism':

> It is an empirical fact that moral judgment in humans is a kind of rational judgment; that is, our moral judgments derive from our rational faculties or capacities.[291]

As its name suggests, this claim is a descriptive claim about how moral judgments are 'produced' – it is claiming that when we make a moral judgment, it is produced by, or flows from, our rational capacities. The second idea he identifies is what he calls 'Conceptual rationalism':

> It is a conceptual truth that a moral requirement is a reason for action.[292]

This is the idea that it is part of the concept of morality that moral requirements are rational requirements: it is true as part of our concept of morality that if one acts immorally it entails that one is also acting irrationally. Nichols thinks that a variety of empirical evidence has implications for both these versions of rationalism. In the following sections I examine whether these are relevant philosophical positions, and then evaluate the success of his arguments about his empirical and conceptual rationalisms.

6.2.1 Are Nichols' rationalisms positions held by philosophers?

Nichols presents the following quotes which he takes to be representative of the ideas of a number of moral rationalists:

---

[287] *Ibid*., p. 65.
[288] *Ibid*.
[289] *Ibid*., p. 66.
[290] *Ibid*., p. 70.
[291] *Ibid*., p. 67.
[292] *Ibid*.

1. "Just as there are rational requirements on thought, there are rational requirements on action, and altruism is one of them…If the requirements of ethics are rational requirements, it follows that the motive for submitting to them must be one which would be contrary to reason to ignore."[293]
2. "The objective badness of pain…is…just the fact that there is reason for anyone capable of viewing the world objectively to want it to stop. The view that values are real is… that our claims about value and about what people have reason to do may be true or false independently of our beliefs and inclinations."[294]
3. "The Kantian approach to moral philosophy is to show that ethics is based on practical reason: that is, that our ethical judgments can be explained in terms of rational standards that apply directly to conduct or to deliberation. Part of the appeal of this approach lies in the way that it avoids certain sources of scepticism that some other approaches meet with inevitably. If ethically good action is simply rational action, we do not need to postulate special ethical properties in the world or faculties in the mind in order to provide ethics with a foundation."[295]
4. "If our concept of rightness is the concept of what we would desire ourselves to do if we were fully rational, where this is a desire for something of the appropriate substantive kind, then it does indeed follow that our moral judgments are expressions of our beliefs about an objective matter of fact."[296]

Nichols writes, that as these passages show, "the consequences are profound and reassuring if moral rationalism is true"[297] but that despite this appeal, the empirical evidence he presents shows that "rationalism is an implausible view."[298] Which then of these quotes supports Nichols' empirical and conceptual rationalism interpretations?

[293] Thomas Nagel, *The possibility of altruism*, p. 3.
[294] Thomas Nagel, *The view from nowhere*, p. 144.
[295] Christine Korsgaard, 'Skepticism about practical reason', p. 311.
[296] Michael Smith, *The moral problem*, p. 185
[297] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment,* p. 67.
[298] *Ibid*.

Quote 1 from Nagel is claiming that there are rational requirements on how we should act – we ought to act ethically because not doing so would be irrational. Nagel is talking directly about morality here, not simply the concept we hold of it, so this quote is not about the conceptual issue that Nichols' conceptual rationalism concerns. Nagel's concern is with *why* we should act ethically; with our "motive for submitting" to moral requirements. Nagel's quote also does not appear to concern how moral judgments are proximally caused – instead it is concerned with the standard of evaluation of those judgments – whether acting in accord with them would be 'rational' or not.

Quote 2, also from Nagel, is stating that the objectivity of moral requirements means that everyone has a reason to act in accord with them, and that these reasons apply regardless of any beliefs or desires to the contrary. Again, neither of empirical rationalism or conceptual rationalism captures this idea of why we should act ethically and the independence of this fact from our beliefs and desires.

Quote 3 from Christine Korsgaard is about showing that ethics can avoid various kinds of moral scepticism by giving it a rational 'foundation' – a basic justification for why we should act ethically. Thus, here again the claim is that moral rationalism is a theory that shows that the reason we should act in accord with morality is because doing so is rational. Quote 4, the final quote that Nichols provides, from Michael Smith's *The Moral Problem*, appears to concern something like Nichols' conceptual rationalism and the implications of it being true. Smith claims that *if* it is true that our concept of rightness (morality) is "what we would desire ourselves to do if we were fully rational" then moral judgments express beliefs about something we believe to be an objective matter of fact.

It is extremely difficult to draw conclusions about moral rationalism in general, as it is a term used in many different contexts for a number of theories or ideas. This variety of theories connects morality to rationality in a range of different ways, and as such there is no single thesis that is denoted by the term 'moral rationalism'. Nichols' interpretation of the various moral rationalist claims he cites seems to miss this point, and further, Nichols' interpretations fail to target perhaps the most philosophically important moral rationalist idea.

The first three of these quotes are all about why we should act in some particular way: they concern rational justification for acting ethically. This suggests that a large part of what moral rationalists are talking about is not Nichols' empirical or conceptual rationalist claims, but the idea that we ought to act ethically because it would be rational to do so or irrational to not do so. This idea we might call 'Justificatory Rationalism':

> Moral requirements are justified on the grounds that it would be irrational not to do as they recommend: we should not act immorally, because doing so is irrational. [299]

This interpretation captures the main elements of the first three quotes that Nichols includes. Additionally, Nichols notes that one of the main motivations for 'moral rationalism' is that it has been seen as one of the most promising ways of securing a kind of objectivity for moral claims.[300] However, he does not discuss why objectivity itself may be desirable, and this is important for understanding moral rationalism. In general, the reason that moral objectivity has been seen as attractive is that objectivity imbues moral claims with a kind of authority or special importance: if moral imperatives are objectively true, then we have a powerful motive to do what they say. So, the focus of many moral rationalists appears to be on why we should act ethically; objectivity, so the argument goes, implies importance or authority, which implies we have reason to act as morality dictates.

The implication of this for Nichols' argument is that it significantly weakens his claims that "rationalism is an implausible view."[301] Firstly, because it is unlikely that most who identify themselves moral rationalists would endorse Nichols' conceptual or empirical rationalisms as fair interpretations of their views on morality. Secondly, because in general, one of the main motivations for moral rationalism is to provide an account of rational justification for following the dictates of morality and none of the

[299] Richard Joyce suggests a similar idea in 'What neuroscience can (and cannot) contribute to metaethics'.
[300] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*. p. 66.
[301] *Ibid.,* p. 67.

considerations or arguments against empirical or conceptual rationalism are applicable to such a justificatory project.

A justificatory rationalism is compatible with just about any account of the proximal mechanisms that produce moral judgment. Whatever the proximal mechanisms involved turn out to be, it has little bearing on whether we would be justified in acting on those judgments. If the proximal mechanisms that produced moral judgments were some kind of intuition, or were simply the expression of an attitude or emotion, we could still rationally evaluate whether acting in accord with such judgments would be justified. Thus, justificatory rationalism appears to be unaffected by the status of empirical rationalism.

Justificatory rationalism is also unaffected by the truth of conceptual rationalism, at least based on any considerations Nichols provides. The fact the concept of morality includes the feature that someone is acting irrationally when they act immorally, is independent of the fact that actually justifies moral judgments. The concepts that people hold do not necessarily track the truth, thus conceptual rationalism and justificatory rationalism can be true or false independently of each other. Further, it is hard to see how any possible descriptive information might have an impact on justificatory rationalism as it is *ex hypothesi* an attempt to base morality on a priori considerations and thus almost no empirical data (other than perhaps if we were to discover that no one existed that met the requirements for being rational) could show that we cannot justify our actions rationally. This is not to say that justificatory rationalism is in any way vindicated by the lack of bearing these empirical issues have on it. For it may turn out that we cannot rationally justify the content of our moral judgments purely by providing an a priori argument – they may in fact be unjustified. The important point is that empirical data or the truth of Nichols' versions of rationalism does not impact on the truth of the justificatory rationalism.

So, Nichols' interpretations of the claims moral rationalism makes are somewhat off target, and thus his conclusion that 'moral rationalism' *in general* is too strong, simply because it misses what is

perhaps the main motivation and idea of moral rationalism. Nevertheless, while his arguments do not generalise to all of moral rationalism, Nichols is still correct that both conceptual and empirical rationalism are claims of interest and importance. In the next sections I examine Nichols' arguments from the psychological evidence to conclusions about empirical and conceptual rationalism, beginning with empirical rationalism.

6.2.2 Empirical rationalism

The idea of empirical rationalism is that moral judgments are the result of a rational mental process. When one makes a novel moral judgment about some issue or event, according to the empirical rationalist the judgment is caused by or generated from, some kind of process of reasoning or rational deliberation. Empirical rationalism holds that moral judgments are always the result of a rational process: it is not simply the view that moral judgments *sometimes* causally flow from rational faculties, for if this was the thesis, then it would be straightforwardly true. Nichols gives the following example that shows this. Sometimes we trust a person's moral views on a range of issues. On this basis, we might accept their testimony that it is morally wrong to buy new furniture made from the wood of old-growth forests and come to hold this view ourselves. Thus it looks like we have reasoned our way to a conclusion that buying such furniture is morally wrong.[302] If cases such as this were all that was required to show empirical rationalism is true, then it would be an easily established, but relatively unimportant thesis.[303] However, the point of empirical rationalism is supposed to be that rational faculties are the "basic font of our moral judgment"[304] – that *all* moral judgments if traced back to the original evaluation that gave rise to them, are ultimately the result of rational deliberative processes. Thus, in the above example, an empirical rationalist would hold that not only was the derivative moral

[302] Shaun Nichols, 'Moral Rationalism and Empirical Immunity', p. 395.
[303] Easy because we would need only to confirm some cases like the above exist, and unimportant because it does not necessarily tell us anything interesting about how novel or 'original' moral judgments get made.
[304] *Ibid*.

judgment arrived at rationally, but that the original judgment made by the person whose moral views we trust, necessarily 'stemmed' from, or was caused by, a rational deliberative process. So, Nichol's empirical rationalism is the thesis that *all* moral judgments originate at their source, from a rational process of deliberation.

Empirical rationalism appears to be a purely descriptive thesis: it simply characterises the proximal mechanisms involved in making novel or original moral judgments as a rational process of deliberation. As such it appears to be a thesis that suitable research in moral psychology could provide evidence for or against. Despite empirical rationalism's apparently purely descriptive nature, philosophy has a long history of theorising about the capacities involved in making moral judgments and their relationship to the nature of morality. One of the reasons for this interest in people's perceptions of having moral obligations is the thought that "many of the deepest issues concerning the nature of morality would be illuminated if we had an adequate account of the nature of moral judgment."[305] One of the issues which Nichols thinks could be resolved is that if empirical rationalism turned out to be true – if human moral judgment were the product of a rational reasoning process, then this fact would provide "justification for thinking that human morality is in fact objective."[306] The reason for this is because if moral judgment derives from our rational faculties, then everyone who has such faculties should arrive at the same moral views. On this point Nichols references Michael Smith who makes the analogy between our convergence on mathematical truths due to a shared capacity for reasoning and morality:

> …something like such a convergence in mathematical practice lies behind our conviction that mathematical claims enjoy a privileged rational status. So why not think that a like convergence in moral practice would show that moral judgments enjoy the same privileged rational status?[307]

[305] *Ibid*., p.398.
[306] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 71
[307] Michael Smith, 'Realism', p. 408.

Based on this analogy, Nichols offers a second characterization of empirical rationalism:

> The psychological capacities underlying moral judgment are, like the psychological capacities underlying mathematical judgment, rational mechanisms.[308]

So, Nichols thinks that an important implication for moral philosophy of such an empirical rationalism being correct is that all rational creatures could converge on agreement about moral claims in the same way they do about mathematical claims. The result of this would be that these moral claims could be seen to have a special status of objective 'correctness'.[309] Thus, while empirical rationalism is itself a purely descriptive claim, Nichols argues that it can be used to argue for a philosophical position with potential philosophical and normative importance: that morality is in an important sense objective.[310]

Nichols argues that evidence from psychology concerning psychopathy shows that empirical rationalism is false. The basic idea of Nichols' argument is simple: people who are classified as psychopaths appear to have intact rational capacities (indeed, in some cases what might be considered exceptionally well-developed rational abilities), and yet they are unable to make and properly comprehend genuine moral judgements. Thus psychopaths appear to have the pre-requisites for making moral judgments according to the empirical moral rationalist, and yet, without exception, they are deficient in their capacity to do so. This, according to Nichols' argument, indicates that rational deliberation is not sufficient for making moral judgments.

However, there are two difficulties with Nichols' argument; a philosophical problem and a problem with the empirical evidence he cites. The philosophical issue is that the question of whether

[308] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment,* p. 69
[309] *Ibid*., p. 72.
[310] There is much needed filling out of what is meant by this, which Nichols doesn't do, which makes it difficult to be sure what his conception of objectivity is (although of course this is itself a difficult task and any provided conception of objectivity is likely to be controversial). The claim that morality is objective appears to be a descriptive claim, but like much of metaethics it is not clear that this is completely true. If I claim that some or all moral claims are objective, then I am (according to many philosophers) saying that it is True with a capital T that you must do some things – and this sounds normative.

psychopaths are able to make moral judgments is dependent upon how we interpret the concept of 'moral judgment'. Second, the evidence concerning psychopaths' capacity for moral judgment that Nichols cites is significantly limited in the scope of its relevance to the broad range of moral judgments that people make, and there are serious empirically backed doubts about the validity of the moral/conventional distinction upon which Nichols' evidence about psychopaths depends. I discuss these two objections in the following sections.

6.2.2.1 What is wrong with psychopaths?

Nichols uses evidence from R. James Blair's work on the performance of psychopaths on the 'moral/conventional task' to support his argument.[311] The moral/conventional task, pioneered in research conducted by Elliot Turiel, aims to empirically explore the development of people's ability to distinguish what Turiel termed 'moral rules' from 'conventional rules'. According to this research, moral rules are characterized as "unconditionally obligatory, generalizable, and impersonal insofar as they stem from concepts of welfare, justice, and rights."[312] In contrast, conventional rules are characterized as "part of constitutive systems and are shared behaviours (uniformities, rules) whose meanings are defined by the constituted system in which they are embedded".[313] Moral rules have the following features:

- Moral rules are not dependent on the authority of any individual or institution: they apply no matter what someone else says.
- Moral rules are generalizable: not only do they apply here and now, they apply to other contexts, including other locations and at other times in history.
- Moral rule transgressions typically involve injustice, harm to specific individuals, or rights violations.

[311] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 76.
[312] E. M. Turiel, 'Morality: Its structure, functions, and vagaries', pp. 169-170.
[313] *Ibid*.

- Moral rule transgressions are considered much more serious than violations of conventional rules.

Conventional rules differ from moral rules in the following ways:

- Conventional rules are rules that facilitate social coordination and organization. Accordingly, they can be suspended or changed by an authority or institution.
- Conventional rules are applicable only in their original contexts. They do not generalize to other locations or times in history.
- Transgressions of conventional rules do not involve injustice, harm, or rights violations.
- Transgressions of conventional rules are less serious than violations of moral rules.[314]

Nichols's examples of moral violations include pulling another person's hair, stealing, pushing another child off a swing, and hitting another person. Examples of conventional rule violations include chewing gum in a class, violations of etiquette (e.g. drinking soup from a bowl), and violations of family rules such as not clearing one's dishes.

Nichols claims that Blair's evidence concerning psychopaths' performance on the moral/conventional task shows that "psychopaths really do have a defective understanding of moral violations."[315] Their defective understanding consists in the fact that although psychopaths generally appear to know right from wrong – they are readily able and willing to pronounce that it is wrong to break into houses, wrong to rob a bank, or wrong to hit others – they do not appear to make any distinction between these kinds of moral violations and the much less serious non-moral violations of rules or conventions. Non-psychopathic individuals make significant distinctions concerning permissibility, seriousness, and authority contingence on the moral/conventional task.[316] Psychopaths do not make any distinctions on these dimensions. Further, the justifications psychopaths give for moral rules typically resemble

[314] List of the characteristics of moral/conventional norms adapted from Daniel Kelly and Stephen Stich *et al*., 'Harm, affect, and the moral/conventional distinction', p. 118.
[315] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 76.
[316] R. J. R. Blair, 'A cognitive developmental approach to morality: investigating the psychopath'.

the justifications non-psychopathic subjects give for conventional rules, and rarely refer to the welfare of victims or the harm done to them as non-psychopaths do when attempting to justify moral norms. As Nichols summarises psychopaths' deficiency, "although there is a sense in which psychopaths do know right from wrong, they don't know (conventional) wrong from (moral) wrong."[317]

Thus, Nichols thinks that we are justified in maintaining that psychopaths do not qualify as making moral judgments, despite their being apparently fully rational. If this is the case, then the argument against empirical rationalism could go through by citing a real world, actual example of individuals who are fully rational and yet do not make moral judgements.

However, the status of psychopaths' moral judgment-making capacities is not clear-cut. Psychopaths do appear to make the same kinds of pronouncements on moral questions as non-psychopathic individuals, but they do not appear to be *motivated* by moral prohibitions in the same way normal people are.[318] Therefore, to settle the question of whether the existence of psychopaths shows that empirical rationalism is false, we must be clear about whether this lack of motivation amounts to the fact that we should not attribute to them the capacity to make 'moral judgments'. Does the lack of ability to be motivated to act on moral pronouncements mean that a moral judgment has not been made?

This appears to be a problem for Nichols' argument against empirical rationalism, for the question appears to depend on the answer to a conceptual question; it is not simply a matter of looking at the evidence. This conceptual question is at the heart of a familiar debate in metaethics about moral motivation. On one side of this debate is the view known as moral motivation internalism (simply 'internalism' hereafter). This is the view that one cannot make a sincere moral judgment without being necessarily motivated, at least to some degree, to act in accordance with that judgment. The opposing view, moral motivation externalism (hereafter 'externalism') holds that one can make a moral

[317] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment,* pp. 76-77.
[318] *Ibid*., p77.

judgment without being motivated to abide by that judgment: there is no necessary connection between moral judgment and motivation. According to externalists, while moral judgments do typically cause people to act, the impact that moral judgments have on individuals' deliberations and actions varies widely, and it is entirely possible to make a moral judgment without being personally motivated by it.[319]

Thus, if we think an internalist concept of 'moral judgment' is the correct conception to use, then psychopaths cannot be said to be making moral judgments, since they do not appear to be appropriately motivated by moral judgments or to fully understand their motivational force, importance, or seriousness. Therefore, if our picture of moral judgment is that of internalism, then psychopaths do present a counter-example to empirical rationalism, for they have full rational capacities but do not make genuine moral judgments. At best they could be said to make 'moral judgments' in the inverted commas sense.[320] However, if the conception of moral judgment we accept is an externalist one, then psychopaths present no difficulty for the empirical rationalist. Under an externalist conception, psychopaths meet the requirements for making genuine moral judgments, as it is not a necessary condition that they be motivated by the moral judgments they make. The externalist concept of moral judgment is *ipso facto* one that does not require motivation as a necessary element.

There is no consensus in metaethics concerning which account of moral motivation is correct or what the implications of differing accounts are.[321] While the debate between internalism and externalism is beyond the scope of this thesis, it is sufficient to note that this debate has resulted in something of an impasse, with each side presenting an array of examples in the hope of triggering intuitions they

[319] For an overview of the motivation internalism debate see Gunnar Björnsson, Caj Strandberg, Ragnar F. Olinder, John Eriksson, Fredrik Björklund, *Motivational internalism: contemporary debates.*
[320] This is a typical response to cases such as psychopaths or the 'rational amoralist', see §6.2.3
[321] Gunnar Björnsson, Caj Strandberg, Ragnar F. Olinder, John Eriksson, Fredrik Björklund, *Motivational internalism: contemporary debates*, pp. 12-16.

feel support their position.[322] The implication of this for Nichols' argument is that it cannot simply assume one account of moral motivation, and without being able to do so, it is hard to see how evidence on psychopaths can be taken to provide a counter-example to the idea of empirical rationalism.

6.2.2.2 The moral/conventional distinction

The second difficulty for Nichols' argument against empirical rationalism is that there are questions about the robustness of the moral/conventional distinction, especially with regards to its use or application for resolving questions about the nature of moral judgment itself. Early work on the moral/conventional distinction focused explicitly on developmental moral psychology and used young children as subjects.[323] The examples of moral and conventional rule transgressions in such studies were of the kind that would be familiar to such children – transgressions such as hair pulling, pushing another child off a swing, or chewing gum in class. Subsequently the research was expanded to examine the responses of a wider range of subjects to the moral/conventional task. However, the range of transgressions used in the experimental task was not similarly expanded. Nearly all of the transgressions used in subsequent studies were of the same schoolyard variety as the early research – the kinds of examples used in the original studies so as to be familiar to very young subjects. The evidence Nichols cites regarding psychopaths' performance on the moral/conventional task also uses these kinds of transgressions, despite the fact that the subjects in question are convicted adult criminals in adult jails.[324] Thus the evidence about the moral/conventional distinction is quite limited in its scope. There is a good chance that convicted criminals will view all of the schoolyard transgressions described as being of similar seriousness. It seems highly plausible that they might find

[322] For example, see R. Francén, 'Moral motivation pluralism' and Michael B. Gill, 'Indeterminacy and variability in meta-ethics'

[323] J. G. Smetana, 'Preschool children's conceptions of transgressions: Effects of varying moral and conventional domain-related attributes' and D. Weston, E. Turiel, 'Act-Rule relations: Children's concepts of social rules'.

[324] R. Blair, 'A cognitive developmental approach to morality: Investigating the psychopath', and R. Blair, L. Jones, F. Clark, and M. Smith, 'The psychopathic individual: A lack of responsiveness to distress cues'.

being caught chewing gum in class and a bit of schoolyard hair-pulling to both be at the less serious end of the scale of transgressions. Compared to serious transgressions such as murder or robbery or rape, the schoolyard-type transgressions are likely to be lumped together at the trivial end of the scale. Given that empirical rationalism is supposed to cover the whole range of moral judgments, Nichols' argument is based on a very limited sample for the general conclusions about empirical rationalism that he reaches.

This difficulty could be remedied by further research that considered a wider range of transgressions on both normal and psychopathic populations. If the same pattern of responses were found – if normal populations make the same distinctions and psychopaths do not – then we could be more confident that the research was representative of the full range of moral judgments. Part of this research has already been done. A study on non-psychopathic individuals was conducted by Kelly et al., who noticed this fact about all of the prior moral/conventional task studies using schoolyard type examples of behaviour.[325] Kelly et al. examined people's performance on the moral/conventional task using a much broader set of transgressions. They found that for the moral transgressions they tested (including, a captain of a cargo ship administering excessive discipline, abuse of military trainees, and cannibalism), most or all of the dimensions considered part of the signature moral response were missing or inconsistent. When more general, adult transgressions are used, the distinctive pattern of responses to the moral/conventional task is not evident. Subjects did not make the distinctions between authority independence, generalizability, or seriousness, which the paradigm predicts. As Kelly et al. focused only on a sample from the general population, there is no further data on psychopaths. However, even were such a study to be done, the Kelly et al. study suggests that there would not be a clear pattern of signature responses in the moral/conventional task that normal populations make to compare them too. They think that if they are correct, then their results suggest that "the moral/conventional task is not a good assay for the existence of a psychologically important

[325] Daniel Kelly, Stephen Stich, Kevin J. Haley, Serena J. Eng, and Daniel Fessler, 'Harm, Affect, and the Moral/Conventional Distinction'.

distinction",[326] and thus we should be cautious about concluding too much about moral judgment based on these possibly questionable, and definitely narrowly sampled research results.

6.2.2.3 Is empirical rationalism supported by the evidence?

Empirical rationalism is the theory that moral judgments are produced by 'rational' mental faculties. It is one of two rationalist ideas that Nichols identifies as being impacted by empirical research in moral psychology. The philosophical importance of empirical rationalism comes from the idea of morality being objectively true in the same way that mathematical claims appear to be objectively true. In principle, it appears to be a question that is tractable to empirical methods, as the thesis simply describes which mental processes are at work when we make moral judgments. However, Nichols' attempt to use evidence concerning psychopaths to show that it is false faces two significant difficulties. The first is that it is not clear that the evidence shows something about moral judgment. Instead the evidence tells us something about moral motivation, what people do once they have made a judgment on some moral issue. This is a problem, as the debate about moral motivation is split into two broad camps: internalists and externalists. Only internalism holds that motivation is necessary to have made a moral judgment, and thus Nichols' argument is only successful if we assume internalism is true. Thus, there is a conceptual issue that requires settling before the evidence can be useful in settling the empirical issue.[327] The second difficulty for Nichols' argument is that the evidence concerning psychopaths makes use of research based on a distinction that requires further research before it is clear whether or not it is robust or psychologically real and important with respect to moral judgments in *general*. At the very least, studies involving psychopaths need to be done which test for more general kinds of moral transgressions. In the next section I evaluate Nichols' claims concerning the second rationalist thesis: conceptual rationalism.

[326] *ibid.*, p. 129.

[327] If indeed the dispute is solvable; it is possible that the concepts involved are not determinate enough for there to be a matter of fact on the issue, as discussed in §6.2.3.

6.2.3 Conceptual Rationalism and moral motivation

According to Nichols, the basic idea of his conceptual rationalism is that it is a conceptual truth that a moral requirement is a reason for action. As previously mentioned, this is one of a class of theories concerning the connection between making a moral judgment and the reasons or motivation to act in accord with that judgment, usually called 'moral internalism' (or just 'internalism').

The core claim of internalism is that:

> If an agent judges they are morally required to φ, then they are at least to some extent motivated to φ. [328]

This will be referred to as the 'simple' version of internalism in what follows. It easy to see why this is an appealing position. We make moral judgments to provide guidance on what we ought to do, so it would be odd if there were no connection that tied the moral judgments made and the causes of actions following those judgments together. The opposing view, that an agent who makes a moral judgment need not be motivated by it, is termed 'externalism'.

However, the simple version of internalism is often taken to be saying too much, as it implies there are no cases where moral judgment and moral motivation can exist without each other. The externalist argues we can countenance 'amoralists' who we can conceive of making a moral judgment, but not being motivated to act accordingly. The internalist response is to argue that when this happens, it can be explained in a number of ways that does not impugn the internalist intuition that moral motivation is inherent to the concept of moral judgment. The explanations given might be that it is because of some psychological malady such as apathy, depression, exhaustion, emotional disturbance or simply a lack of rationality. In response to this, the simple version of internalism is

[328] Gunnar Björnsson, Caj Strandberg, Ragnar F. Olinder, John Eriksson, Fredrik Björklund, *Motivational internalism: contemporary debates*, p. 7. Note that the wording has been changed to be in line with that used by Michael Smith's Practicality Requirement below, but nothing hinges on this for the present purposes.

usually refined to account for these potential conditions by adding a caveat that the motivation need not follow judgment if certain conditions obtain:

> If an agent judges they are morally required to φ, then they are at least to some extent motivated to φ, as long as not condition C.

The condition C describes the type of defect the amoralist suffers from that allows us to explain their lack of moral motivation. In surveying the literature, Bjornson et al. identified three broad kinds of conditions that defenders of internalism have offered as a way of accounting for cases where motivation is lacking without these cases constituting 'amoralist' counterexamples:[329]

1. Psychologically abnormal: Normal psychological functioning is required for motivation to be expected as a result of moral judgment. Thus, moral judgments will be motivating except in cases where the agent is not in a psychologically normal state: where they are depressed, apathetic, exhausted, emotionally disturbed, or in some other disordered mental condition.
2. Morally imperceptive: If an agent does not truly perceive the moral nature of a judgment, they may appear to make moral judgments but are really 'going through the motions' and thus a lack of associated motivation can be explained.
3. Practically irrational: The types of conditions outlined in 1. above are mitigating factors because in some sense they remove from the agent in question the fullness of their usual deliberative abilities and rational control over actions and desires. That is, we expect moral judgments to imply motivation to some degree when an agent is rational. A lack of moral motivation following a moral judgment can be explained if we know the agent is not rational.

[329] *Ibid.*, p. 7.

The variety of internalism that Nichols identifies as his target as part of his argument against conceptual rationalism is of this last kind, in particular a version defended by Michael Smith.[330] This feature of this form of internalism Smith terms the 'Practicality Requirement'[331]:

> If an agent judges that it is right for her to φ in circumstances C, then either she is motivated to φ in C or she is practically irrational.[332]

According to the Practicality Requirement, if an agent makes a moral judgment *and* they are rational, they will be motivated to act in accordance with the moral judgement. Conversely, the Practicality Requirement implies that if the agent were irrational, it would be no contradiction to the truth of conceptual rationalism if they were to make a moral judgment but did not act in accord with that judgment.

Nichols argues that the concept people hold of a psychopath provides us with a counter-example to the Practicality Requirement. Nichols holds that because psychopaths are taken to be rational, they cannot fall under the group that are 'practically irrational' and yet, for psychopaths, moral judgments and moral motivation do not go hand in hand. Nichols argues therefore that the concept of a psychopath is that they are rational, they know what is right or wrong, but simply do not care and thus are not influenced by those judgments.

However, instead of going down the usual path of conceptual analysis to attempt to establish this conclusion about the features of our concept of a psychopath (which as noted earlier, he argues readily devolves into a stalemate), Nichols attempts to use empirical evidence to settle the debate.

---

[330] Nichols is explicit his target is Smith's Internalism and the Practicality Requirement, see *Sentimental rules: On the natural foundation of moral judgment*, p. 71.

[331] Nichols writes "For current purposes, the crucial feature of conceptual rationalism is its account of the link between moral judgment and motivation. Smith maintains that conceptual rationalism entails the Practicality Requirement, according to which 'It is supposed to be a conceptual truth that agents who make moral judgments are motivated accordingly, at least absent weakness of the will and the like' (Smith 1994, 66)." Note that this is not Smith's definition of the Practicality Requirement which Smith explicitly gives on p. 61 and labels as such on p. 62. Instead what Nichols identifies it is part of a comment Smith makes about the Practicality Requirement on p. 66. Smith's actual definition is as provided above.

[332] Michael Smith, *The moral problem*, p. 61.

Nichols carried out surveys of undergraduate students by presenting them with vignettes about a fictional psychopath 'John' who says he knows that it is wrong to hurt other people but does not care if he does things that are wrong. Survey respondents are then asked "Does John really understand that hurting others is morally wrong?" with the intention of establishing whether respondents' concepts of moral judgment includes that John could be making a real moral judgment without being motivated by it. The full description and question given to survey participants is as follows[333]:

> **Description**: John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt and indeed killed other people when he has wanted to steal their money. He says that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong.
>
> **Question**: Does John really understand that hurting others is morally wrong?

The responses to the survey were that nearly 85% answered that John does understand that hurting others is morally wrong. Nichols concludes on the basis of this result that "the common conception of psychopaths is precisely that they really do know the difference between right and wrong, but they do not care about doing what's right"[334] and that "[c]ontrary to the conceptual rationalist claims, psychopaths are commonly regarded as rational individuals who really make moral judgments but are not motivated by them."[335]

There are a number of comments to make about Nichols' argument here. Firstly, as Daniel Eggers has noted, Nichols' questions do not focus strongly on moral motivation.[336] The description Nichols' survey gives of John does not make it explicit that John lacks moral motivation. The only indication given is John's self-report that "he just doesn't care if he does things that are wrong", and it is not

[333] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 74.
[334] *Ibid*.
[335] Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*, p. 82.
[336] Daniel Eggers, 'Unconditional motivational internalism and Hume's lesson, p. 100.

clear that the concepts of 'not caring' that things are wrong and 'lacking all moral motivation' are the same thing; it certainly is not an explicit connection.

The second thing to note is that while Nichols' evidence may provide *prima facie* support for rejecting the simple version of internalism, it is less clear that his argument can apply in the way he claims to Smith's conditionalized version: to the Practicality Requirement. Nichols has targeted his view at Smith's version because it is apparently about rationality, something that is purportedly not a deficiency of psychopaths.[337] However, as pointed out by Caj Strandberg and Fredrik Björklund,[338] Nichols' description does not mention practical rationality, or in fact, rationality at all. The only comment that comes close is that John is of 'normal intelligence'. There is nothing in the description that directly addresses the question of whether a psychopath could be considered 'practically irrational' and one can suffer from any number of rational defects while still being of normal intelligence. Thus, it is entirely possible that participants responded that John did understand that hurting others was wrong, but with the idea in mind that he is also not a rational person (or is suffering from some condition that would constitute practical irrationality). The above two considerations mean Nichols argument in its current form is not successful; the evidence Nichols gives does not show Smith's version of internalism to be false. However, this is not necessarily fatal, it would be possible to modify the description of John to explicitly check that respondents thought he was not practically irrational.

Strandberg and Björklund set out to do this by attempting to replicate Nichols results but also refining them with explicit checks so as to avoid these objections. They undertook surveys which describe various actors in different mental conditions and attempt to more carefully formulate their scenarios.

[337] Nichols examines the possibility that psychopaths suffer from a systematic defect in their faculty of reason but rejects the possibility, see Shaun Nichols, *Sentimental rules: On the natural foundation of moral judgment*., pp. 78-81.
[338] Caj Strandberg and Fredrik Björklund, 'Is moral internalism supported by folk intuitions?'

They began by providing a simple scenario to test people's responses about moral motivation, then varied it by adding in additional pieces of information. The simple scenario was:

> **Simple**: Anna is watching a TV programme about a famine in Sudan. In the TV programme, it is shown how the starving are suffering and desperately looking for food. At the same time, Anna is not motivated at all, not to any extent, to give any money to those who are starving.
>
> **Question**: Could it be the case that Anna thinks she is morally required to give some of her money to the starving even if she not motivated at all to do so?

The variations on this initial scenario included stipulating that Anna is mentally healthy and normal or that she was apathetic, deeply depressed or a psychopath, and her lack of motivation in these later cases was due to the condition described in each. Contrary to Nichols' results, a majority of the participants responded in ways indicating that both simple and conditional versions of internalism are not supported by the linguistic intuitions of the sampled population. For the Simple case, a majority of respondents affirm that Anna is making a moral judgment despite not being motivated to act accordingly. In the cases where it is also specified that Anna is mentally 'normal functioning', a majority of respondents reported it was possible that Anna made a moral judgment without motivation. The cases where Anna is described as apathetic or deeply depressed also had majorities of respondents affirming an externalist understanding of the concept of moral judgment where motivation did not necessarily follow. Interestingly in the Psychopath case only a minority of respondents said it was possible Anna made a moral judgment without being motivated by it. This is the opposite result to that reported in Nichols' survey (and the only result that might support an internalist conception of moral motivation which would run counter to Nichols' argument).

Thus, according to Strandberg and Bjorklund's study a majority of the participants regarded it as entirely possible that someone can make a moral judgment without being motivated by that judgment. They conclude that this is "contrary to what they [respondents] should be expected to do

on the assumption that any of these internalist claims were correct"[339] and that simple and conditional internalism do not appear to be supported by the folk intuitions of the study participants. The modifications to the surveys that Strandberg and Bjorklund made address the criticisms raised earlier against Nichols' survey, while the results are contrary to those of Nichols. However, there are other difficulties for Nichols and Strandberg and Bjorklund's conclusions. The difficulties concern the forcefulness of their claims, when it is not clear that the evidence actually supports such strong conclusions. The reason for this is the manner in which the empirical results are interpreted as supporting or not supporting their conclusions. The nature of the support that a majority of survey respondents provides for a philosophical conclusion is left unexamined and looks likely to be a significant weakness for their arguments.

In both Nichols and Strandberg and Bjorklund's studies, agreement by a majority of anywhere from 60% to 84% of respondents in a survey is taken as more or less conclusive that the conceptual feature identified is a platitude about that concept. So even in the case with the highest majority, where 84% of Nichols' respondents said that John the psychopath made moral judgments but was not motivated,[340] 16% of respondents held the contrary view. In Strandberg & Bjorklund's experiments the results were even less conclusive. The majorities were Simple: 76%, Normal functioning: 79%, Apathy: 60%, Depression: 79%. The least conclusive case is precisely that which Nichols argument is based on – Psychopath, where 42% of the respondents had externalist intuitions about motivation, which means presumably that the remaining 58% of respondents had internalist intuitions. This is not a result that can be interpreted as providing clear support for either camp. Focusing on the actual figures for different responses in the surveys shows that while there may be a majority in each case, this does not constitute a consensus, or anything close to it, about people's intuitions about cases of moral judgment and corresponding moral motivation.

[339] *Ibid.*, p. 335.
[340] Shaun Nichols, 'Is it irrational to be amoral? How psychopaths threaten moral rationalism', p. 289.

What then should we make of the fact that for some it is a platitude that an individual can be judged as rational and make a moral judgment without it being motivating, while for others (a minority in most of the tested cases, but still large proportions of the respondents) this is not a possibility? The way in which Nichols and Strandberg and Björklund appear to be interpreting the results is that if there is any majority (although if that is indeed the cut-off point it is not made explicit anywhere), then the result is taken as wholesale and decisive evidence supporting only one interpretation of the concept. But interpreting the results in this way misconstrues the results of the survey. For what this kind of evidence is really telling us about respondents' intuitions is something quite different to a decisive verdict on a concept. A more accurate interpretation would be one of two possibilities.

The first possibility is that different respondents may have interpreted the question in different ways (or comprehended the content and facts of the situation differently). For example, in Nichols' questionnaire, some may have had the expectation that John is practically rational based on the comment that "he is of normal intelligence" whereas others may have not interpreted this as any kind of commentary on his practical rationality. If this is the case, then the results of the surveys do not reflect the participants' intuitions on the intended question.

Alternatively, where the interpretation of the scenario by participants is relatively uniform and the questions are well elaborated and explicit about features of concepts they are examining, different proportions of respondents having differing intuitions may simply be due to not sharing the same platitudes about the concept in question. What the surveys discover, if this is the case is, that there is no single interpretation or meaning of the concepts involved – respondents have differing working analyses of that concept.

Where 42% answer in one way, and 58% another, such as in Strandberg and Bjorklund's psychopath case, all we can actually know about this sample is that there is disagreement with marginally under half categorising it into one of the two options, and remainder into the other option of those presented in the survey. Even if we did have a representative sample of everyone who used these

concepts, a mixed result would still leave us with a question of interpretation that must be addressed. This problem of reconciling different intuitions in conceptual analysis is not a new one; essentially, we are back to the same situation as that faced in traditional conceptual analysis where different philosophers do not share intuitions.[341] In such situations of conflicting intuitions there is no established or agreed upon procedure for resolution, beyond focusing on the arguments for each position, or trying again with different thought experiments in the hope of eliciting more consistent or persuasive intuitions.[342]

For now, I will simply conclude that much more would need to be said for Nichols, Strandberg and Bjorklund to have been successful in establishing arguments that would or should be accepted by moral internalists or externalists as settling the issue one way or the other. Especially, given the known lack of diversity in the samples in the studies presented (which surveyed groups of undergraduates at the respective universities of Nichols, Strandberg and Bjorklund), we should refrain from drawing conclusions based on these results as indicative of anything about the concepts themselves. At most, their surveys establish that there is some variability of concepts of moral judgment and the modality of the link between moral judgment and moral motivation within philosophically untrained undergraduates. This variability appears to be sensitive to a number of influencing factors, as highlighted by Strandberg and Bjorklund, but not limited to those they identify. The variability in

[341] There are some interesting differences in analysing the differences between the traditional first-person approach to conceptual analysis and a third-person approach. For a comparison of the merits see Kirk Ludwig, 'The epistemology of thought experiments: first person versus third person approaches.' Ludwig examines among other factors, "how…[survey respondents] understand the task, their background beliefs, empirical and nonempirical, how they think what they say will be taken, loose analogies they may draw with other sorts of situations, how they understand the scenario, whether they pay adequate attention to relevant details, whether they think clearly and hard enough to see what to say in response to the kind of question asked, assuming they understand it correctly, how they think that their interlocutor will (or interlocutors generally would) understand what they say or more generally what they would be trying to convey by what they say or how they respond, as well as perhaps various shortcuts or rules of thumb in reasoning, or plain mistakes", p. 144.

[342] It has also been argued that intuitions of these sorts are not actually the basis of conceptual analysis – at least not in the sense of being evidence for or against a particular interpretation of a concept. Instead, while conceptual analysis may make use of hypothetical cases or thought experiments and intuitive reactions to these scenarios, it is the arguments that philosophers make about these intuitions that should more rightly be treated as the evidence for the claims in question. See Max Deutsch, *The myth of the intuitive: Experimental philosophy and philosophical method* for a defense of this theory. Such a thesis would make the present empirical results interesting, but not helpful in settling matters.

intuitions is interesting, and it may be that the irresolvability of the issues around moral motivation is partly due to this variation and different understandings – the concepts involved may simply not be determinate enough, in the sense of everyone sharing the same meanings.

6.3 Adina Roskies and moral motivation internalism

Adina Roskies has advanced an argument that is similar to that made by Nichols. Roskies however has focused expressly on motivational internalism rather than conceptual rationalism. Roskies contends that motivation internalism can be shown to be false based on empirical considerations. She argues that certain brain-damaged patients constitute walking counterexamples to an internalist conception of moral judgment and further that her argument "stands as an example of how empirical evidence can be relevantly brought to bear on a philosophical question typically viewed to be *a priori.*"[343]

Roskies formulation of motivation internalism that is the target of her argument is as follows:

> If an agent believes that it is right to Φ in circumstances C, then he is motivated to Φ in C[344]

This view is what she calls the 'substantive internalist thesis'. She argues that this substantive internalist claim is false and can be shown to be so due to empirical evidence. Roskies' walking counterexamples which show the substantive internalist thesis to be false are patients who have suffered damage to the ventromedial prefrontal cortex area of the brain later in life (hereafter VM patients). These VM patients have what neuroscientist Antonio Damasio has termed 'acquired sociopathy'. They are able to make appropriate moral judgments when queried and appear to make the same judgments as normal people. Despite this capability to make moral judgments, their ability

[343] Adina Roskies, 'Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy"', p. 52.
[344] *Ibid.,* p. 55.

to act effectively on these judgments is impaired. They report and display a lack (or reduced intensity) of affect when faced with moral situations that reliably elicit emotions in normal subjects. [345]

Roskies interprets the character of VM patients' acquired sociopathy to have two features that are important for her argument. Firstly, their mastery of the concept of moral judgment and ability to employ it normally in making moral judgments means they are suitable test subjects for investigating the connection between moral judgment and motivation. The second relevant feature that Roskies argues VM patients display is that they show a striking lack of motivation following many of the moral judgments that they make. Together these two features mean that VM patients constitute a counterexample to the idea that motivation is necessarily part of the concept of moral judgment, as they are agents who make such judgments, but are not motivated by them. Note that VM patients need not be lacking motivation following all the moral judgments they make, the conceptual possibility could be demonstrated by VM patients only sometimes lacking motivation following a moral judgment.

In defence of the first of these characteristics, that VM patients should count as making real moral judgments, Roskies argues there is no evidence or reason to think that VM patients who were previously competent at moral judgment subsequently lose this capability. Tests and interviews with VM patients show no sign that they lose their mastery of moral language or ability to reason about moral ideas.[346] There is also no evidence to show that VMPFC lesions affect declarative knowledge of any kind, including knowledge about moral norms and categories of moral behaviour. Unlike psychopaths, VM patients can be presumed to have functioned at a normal level of moral competency prior to their injuries. To avoid being entirely ad hoc, any argument that VM patients lack mastery of moral knowledge will have to explain why and how they happen to lack this knowledge despite there

[345] General characterisation of VM patients adapted from *Ibid.,* pp. 56-57.

[346] For example, VM patients appear to respond with normal competency in structured interviews that attempt to assess the level of moral competency on Kohlberg's moral reasoning scale. See J. L. Saver & A. R. Damasio, 'Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage'

being no apparent deficit or independent reason to expect a deficit. Roskies highlights this analogously: it might be plausible to maintain that a congenitally blind person never has full knowledge of colour terms, but it would be implausible to argue that a newly blind (or person closing their eyes) suddenly loses knowledge of the meaning of colour terms if they happen continue to use those terms in the same way once they can no longer see.[347]

Additionally, to argue that VM patients only make moral judgments in some inverted commas sense would also appear to be groundless. Unlike the prototypical 'amoralist' of philosophical literature, with VM patients there is no reason to believe they wish to seem as though they make moral judgments (either consciously or otherwise) while in reality attempting to hide their real intentions so as to appear as a genuinely morally concerned individual. VM patients respond to experimenters similarly about their moral and non-moral beliefs without apparent incongruity or apparent intention to deceive listeners about the real nature of their character. These considerations lead Roskies to be confident in the view that the ability of VM patients to make moral judgments is undisturbed.

The above assessment appears to be persuasive; at least when applied to abstract reasoning about moral situations of the kinds involved in studies of VM patients' moral judgment making abilities. However, importantly for Roskies' argument, the moral judgments that VM patients make, also need to be tested for accompanying motivation. Jeanette Kennett and Cordelia Fine have noted that in the studies of VM patients that Roskies uses as evidence, the scenarios presented to subjects are all moral situations that are both hypothetical and are posed in the third person.[348] The cited studies require that the subject respond to moral dilemmas that involve making judgments about how the hypothetical participants in a moral dilemma ought to act. They contrast this with what they term first-personal 'in situ' moral reasoning – where subjects would make judgments about how they themselves should act in a given moral dilemma.

---

[347] *Ibid.,* p. 60.

[348] Jeanette Kennett, Cordelia Fine, 'Internalism and the evidence from psychopaths and "acquired sociopaths"', pp 181-183, and 'Could there be an empirical test for internalism?', pp. 220-224.

Kennett and Fine argue that these hypothetical third-personal scenarios are not a good test for whether VM patients lack motivation following moral judgment. The reason for this is that such third-personal scenarios do not constitute a situation in which the subject of the interview are ever themselves be motivated to act. For moral judgments of these kinds, in these circumstances, the only possible actors in the scenarios are third parties who also happen to be hypothetical. Thus, Kennett and Fine argue that the evidence Roskies presents about VM patients' ability to make moral judgments, focuses on moral judgments of a kind that could not be used to disconfirm moral internalism. The kind of moral judgment used to test for moral motivation must be one where the moral judgment is made by the subject about their own current situation and actions; it must be a first-personal in situ moral judgment.[349]

Roskies and Michael Smith both separately complain that this restriction to first person in situ judgments mis-understands the nature of moral requirements and looks unnecessarily ad hoc.[350] They interpret the conclusion of Kennett and Fine's argument as a view on the nature of internalism (that it is only applicable to first personal in situ circumstances) and as a restriction on when we should expect moral motivation to follow moral judgment according to motivation internalism.[351] Smith and Roskies argue that moral requirements are by their nature applicable and motivating to all parties at all times; they do not stop being applicable or motivating simply because the judging agent is not themselves involved in a situation requiring an immediate moral choice and action. But this is a misinterpretation of what Kennett and Fine are arguing; their requirement that the relevant kind of moral judgments are first person in situ judgments is solely for the purposes of empirically verifying whether VM patients constitute a counterexample to motivational internalism. Kennett and Fine are

[349] Jeanette Kennett, Cordelia Fine, 'Internalism and the evidence from psychopaths and "acquired sociopaths"', p. 182.
[350] Michael Smith, 'The truth about internalism', pp. 208-210, and Adina Roskies 'Internalism and the evidence from pathology', pp. 194-195.
[351] Smith is quite explicit that this is his interpretation of their argument. He writes "the conclusion of Kennett and Fine's argument…is that we should restrict internalism to in situ judgments", *Ibid.,* p. 209, and "The version of internalism that Kennett and Fine propose…can be stated as follows: Other things being equal, if an agent makes the in situ judgment that she ought to ϕ in circumstances C—that is, if she judges that she ought to ϕ in circumstances C, believing herself to be in those circumstances—then she is motivated to ϕ.", *Ibid.*, p. 208.

not attempting to define a new more restricted version of internalism that their argument will apply to. Instead they are contending that for empirical tests of motivation internalism to be valid, the tests must use first personal in situ moral judgments instead of third personal hypothetical judgments. Thus, the research Roskies cites is not sufficient to show VM patients are making moral judgments that could be tested for motivation. This is not to say however, that such tests would be impossible; just that the such research has not yet been carried out.

For the second consideration important for her argument, that VM patients are not motivated by moral judgments, Roskies provides two kinds of evidence: reports of the behaviour VM patients from case histories and the results of Skin-conductance response tests to 'ethically charged' situations. The reports of VM patient behaviour are limited, coming from descriptions of Phineas Gage and the VM patient referred to in the literature as 'EVR'. Phineas Gage is a commonly mentioned figure in psychology and neuroscience, famous for surviving a severe brain injury from an iron tamping rod being driven through his head and for having notable personality and behavioural changes as a result of this accident. His case was one of the first to suggest that damage to specific parts of the brain might differentially affect different capabilities or personality traits.

However, there are significant challenges to using Gage as evidence about the nature of VM patients. At the time of his injury, reports in medical journals mostly focused on his surprising survival rather than on the impacts of the injury on his behaviour. The descriptions by his physician of his personality and behavioural changes were limited in length to a few hundred words, and have since been re-interpreted and in many cases exaggerated to reflect varying theoretical assumptions of neurology depending on the theory in question.[352] Further, the characterisation of Gage's brain injury as being a ventromedial prefrontal cortex lesion is problematic- both due to limitations of the physical evidence (only his damaged skull is available for analysis) and the reports from the time of his injury in the mid-

[352] For a critical analysis of how Gage has been presented and discussed in neuroscientific literature see Zbigniew Kotowicz, 'The strange case of Phineas Gage'.

1800s. Modern attempts to pinpoint the location of the damage indicate trauma was likely to be limited to the left frontal lobe,[353] while others indicate that the effect to other areas of the brain due to network connectedness was severe and widespread in addition to the trauma to the left frontal lobe.[354] These considerations make Gage an unreliable source of evidence as a VM patient; it is not clear that his brain injury was limited just to the ventromedial prefrontal cortex and our knowledge of his resulting behavioural impairment and case history is both limited and controversial.

The evidence provided for EVR's lack of moral motivation is more reliable, however it is still quite limited in quantity and the interpretation is contentious. Roskies cites Damasio's description of EVR's behaviour that he "…entered disastrous business ventures (one of which led to predictable bankruptcy), and was divorced twice (the second marriage, which was to a prostitute, only lasted 6 months). He has been unable to hold any paying job since the time of the surgery, and his plans for future activity are defective."[355] From this Roskies concludes that "Clinical histories and observation suggest that VM patients are impaired in their ability to act effectively in many moral situations."[356] In response to criticisms about the limited nature of this evidence,[357] Roskies cites a number of studies, but these do not address the motivational aspect of the evidence (that VM patients act immorally despite appearing to be competent moral judgers or are unmotivated by their judgments).[358]

[353] See Peter Ratiu, Ion-Florin Talos, , 'The tale of Phineas Gage, digitally remastered'
[354] See J. D. Van Horn, A. Irimia, C. M. Torgerson, M. C. Chambers, R. Kikinis, and A. W Toga, 'Mapping connectivity damage in the case of Phineas Gage'
[355] Adina Roskies, 'Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy"', p. 56
[356] *Ibid*., p. 57.
[357] See Jeanette Kennett, Cordelia Fine, 'Internalism and the evidence from psychopaths and "acquired sociopaths"', pp. 182-186.
[358] Roskies argues "Kennett and Fine are mistaken in claiming that my evidence comes entirely from the case study of EVR. This is the most detailed case study available in the literature, but similar studies have been carried out with other VM patients, with a similar profile of results (Damasio, Tranel, & Damasio, 1990 ['Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli']; Adolphs & Hauser, personal communication). A new study of seven VM patients (Young, Cushman, Adolphs, Tranel, & Hauser, 2006['Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence']) also shows that these patients' judgments of moral culpability and intention mirror those of normals, in almost every way." However, these studies do not address the motivational aspect or immoral behaviour of VM patients: the paper by Young, Cushman, Adolphs, Tranel, & Hauser discusses the contribution that affect makes to the decision-making process in VM patients; not the motivational state after such decisions have been made. The paper by Damasio, Tranel, & Damasio describes subjects who develop

The second kind of evidence Roskies proposes, to show VM patients lack of motivation following moral judgments, is their atypical Skin-Conductance Responses (SCR) to "ethically charged" scenarios presented to them during studies. Roskies notes that normal subjects produce an SCR response to value-laden stimuli, whereas VM patients do not. She asserts that while SCRs are usually taken as a measure of general physiological arousal, for the purposes of her argument the presence of a measurable SCR can be taken as evidence of motivation, and the lack of an SCR to be indicative of no motivation.[359] This argument however has been criticised for the cited studies not actually testing "ethically charged" scenarios as the stimuli for SCRs[360] and further that contrary to what Roskies claims, SCRs are not a good indicator of moral motivation.[361]

Roskies ultimately accepts the deficiencies of the presently available research, both in relation to the requirement that the relevant moral judgments for an empirical test of internalism must be first-personal and the requirement that SCRs be first established as a reliable indicator of moral motivation. She concludes that her argument could still succeed however, if research was carried with a focus on answering the philosophical dispute between motivational internalists and externalists: "the ideal experiments to resolve this question would (1) test the hypothesis that SCRs are a reliable indication of moral motivation in normals, and (2) measure SCRs in VM patients and normal simultaneously as they make both first- and third-person moral judgments."[362]

There is reason to be sceptical however of this more limited conclusion due to more recent research that does contradict some of Roskies assumptions about the nature of VM patients moral reasoning

---

defects in decision-making and planning that are shown in abnormal social conduct. The conduct often has negative personal consequences for the subjects but is not clearly categorised as moral or immoral.

[359] She argues "this simplification is warranted: the presence of the SCR is reliably correlated with cases in which action is consistent with judgment, and its absence is correlated with occasions in which the VM patient fails to act in accord with his or her judgments. Thus, the SCR is a reliable indicator of motivation for action" Roskies, 'Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy"' p. 57, see also Roskies' endnotes, 9, 10, 11, and 12.

[360] Jeanette Kennett, Cordelia Fine, 'Internalism and the evidence from psychopaths and "acquired sociopaths"', p. 187, and 'Could there be an empirical test for internalism?' p. 221.

[361] Jeanette Kennett, Cordelia Fine, 'Internalism and the evidence from psychopaths and "acquired sociopaths", p.188.

[362] Adina Roskies, 'Internalism and the evidence from pathology' p. 201.

capabilities. There appears to be significant differences between the moral judgments that VM patients make about themselves and their own actions when compared to hypothetical judgments about others. For example, in one study VM patients were presented with 'trolley problems', which require the chooser to either personally push a man off a bridge to stop a trolley from colliding with a group of people or refrain from pushing the man in front of the trolley resulting in the death of the group. Contemplating this choice, VM patients did not show the usual signs of emotional reaction when compared to controls and were likely to decide to push the man off the bridge in contrast to control subjects who were unlikely to do so.[363] VM patients also tend to make decisions based on different and less relevant criteria when making choices for themselves,[364] and tend to look less far into the future, compared to controls.[365] These kinds of deficiencies in judgment and decision making are better able to explain the poor social outcomes in VM patients' lives than the lack of moral motivation that Roskies attributes to them.

Additionally, the current understanding of the mechanisms causing poor performance in VM patients' personal moral decision making, is that there is a lack of emotional input from the damaged ventromedial prefrontal cortex into the reasoning and judgment process that would normally be present. In normal subjects this emotional component attaches itself to various decision options and adds weight to some and detracts from others, prior to a judgment being made. VM patients' absence of this emotional processing as part of their moral judgments and social decision making in their own lives results in their abnormal moral judgments and poor social choices.[366]

[363] Giovanna Moretto et al., 'A Psychophysiological Investigation of Moral Judgment after Ventromedial Prefrontal Damage'.
[364] L. K. Fellows, 'Deciding how to decide: Ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making'.
[365] L. K. Fellows, M. J. Farah, 'Dissociable elements of human foresight: A role for the ventromedial frontal lobes in framing the future, but not in discounting future rewards.'
[366] Neil R. Carlson, *Physiology of Behaviour,* 11th ed., pp 368-370.

6.4 Social psychology and empirically based arguments against virtue ethics

John Doris and Stephen Stich have argued that a number of the findings of psychology have important implications for virtue ethics.[367] Virtue ethics is often considered to be one of the three major approaches in normative ethics, with its basic premise being that instead of focusing on which action is right or good, or what obligations or duties people have, ethics should focus on what sort of person one should be. Therefore, the central question of virtue ethics is 'what traits of character make one a good person?' Virtues or character traits, as used in virtue ethics, are taken by Doris and Stich to be "settled patterns of motive, emotion, and reasoning that lead us to call someone a person of a certain sort (courageous, generous, moderate, just, etc.)"[368] They also write that according to the virtue ethics tradition, "virtues are supposed to be robust traits; if a person has a robust trait, she can be confidently (although perhaps not with absolute certainty) expected to display trait-relevant behaviour across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behaviour."[369]

Doris and Stich's argue that a large body of research in psychology conflicts with claims that virtue ethics makes. The type of research they cite is termed 'situationism' in psychology, which holds that the behaviour of people is extraordinarily sensitive to the situations in which they find themselves. Doris and Stich think there is an extensive body of evidence that supports their argument, and they cite the following findings as representative:

- Mathews and Canon found subjects were five times more likely to help an apparently injured man who had dropped some books when ambient noise was at normal levels than when a power lawnmower was running nearby (80 per cent v. 15 per cent).

[367] John Doris, Stephen Stich, 'As a matter of fact: Empirical perspectives on ethics'.
[368] *ibid.*, p. 117.
[369] *ibid,* pp. 117-118.

- Darley and Batson report that passers-by who were not in a hurry, were six times more likely to help an unfortunate who appeared to be in significant distress than were passers-by who were in a hurry (63 per cent v. 10 per cent).
- Isen and Levin discovered that people who had just found a dime were twenty-two times more likely to help a woman who had dropped some papers than those who did not find a dime (88 per cent v. 4 per cent).
- Milgram found that subjects would repeatedly 'punish' a screaming 'victim' with realistic (but simulated) electric shocks at the polite request of an experimenter.
- Haney et al. describe how college students role-playing in a simulated prison rapidly descended to *Lord of the Flies* barbarism.[370]

Doris and Stich's argument, based on these findings and others like them, is as follows. Virtue ethics assumes that people have robust, reliable dispositions or character traits. If people have these robust character traits, we should expect them to behave consistently and dependably in the manner associated with the specific traits: honest people should behave in a reliably truthful and fair manner, compassionate people in a reliably sympathetic and humane way, and so on for all the virtues. There is however, a large body of research, exemplified by the studies above, which indicates that 'insubstantial' situational factors powerfully influence how people actually act. Because people are so easily induced by various situational factors to act in ways which are contrary to acting virtuously, this research shows that people do not generally have the robust character traits of the kind assumed by virtue ethics. As Doris and Stich state, such research shows that "behaviour is not typically structured by the robust traits that figure centrally in virtue theoretic moral psychology."[371] Virtues of the kind virtue ethics deals with will not be realised, as psychological data has shown that people do not have robust character traits of the kind required. Doris and Stich therefore conclude, "it looks as though

[370] Summarised from *ibid,* p. 118.
[371] *ibid.,* p. 119. By "virtue theoretic moral psychology" Doris and Stich mean the account of moral psychology that virtue ethics assumes or implicitly accepts.

attribution of robust traits like virtues may very well be unwarranted in most instances, programmes of moral education aimed at inculcating virtues may very well be futile, and modes of ethical reflection focusing moral aspirations on the cultivation of virtue may very well be misguided."[372]

6.4.1 Normative and descriptive claims

The first and most obvious reply to Doris and Stich, is to point out that virtue ethics is simply not in the business of making claims about what people are like. Instead, virtue ethics is concerned with how people *ought* to be. That is, it is a normative theory: it is concerned with what character traits we *should* have, what virtues would be best or most appropriate to possess, what kind of person we ought to be. Therefore, it may appear to be a misguided criticism to claim that virtue ethics makes empirical claims that are shown to be false by psychological research, as virtue ethics simply is not in the business of making descriptive claims about what people are like. The claims of virtue ethics are therefore not open to the criticism that they are not in accord with some set of descriptive psychological findings. Doris and Stich appear to avoid mentioning this distinction between the normative and the descriptive approaches of ethics and psychology in their discussion. For example they write that "virtue ethics is marked by a particular interest in moral psychology, an interest in the cognitive, affective, and emotional patterns that are associated with the attribution of character traits" [373] and that "This interest looks to be an empirical interest, and it's natural to ask how successfully virtue ethics addresses it."[374] But of course, if virtue ethics is a normative theory of the kind that has traditionally occupied philosophical ethics, it cannot simply have an 'interest' in moral psychology. Its concern with moral psychology must be much more specific: it has an 'interest' only insofar as it makes normative recommendations. It is concerned with 'ought' statements saying which traits of character best to possess, which virtues to attempt to educate in others, what kind of person

[372] *ibid*, pp. 119-120.
[373] *ibid.,* p. 117.
[374] *ibid.*, p. 117.

to aim to be and so on. And recommendations or ought statements are simply not the kind of thing to be discovered empirically.

The nature of Doris and Stich's argument therefore needs clarification to show why it is not vulnerable to such an objection (although they themselves provide no discussion in the following terms). What Doris and Stich must be claiming, if their argument is to succeed, is of the form of an 'ought implies can' statement. That is, they are claiming that people *cannot* (robustly, consistently), instantiate the virtues prescribed by virtue ethics, and therefore it is not the case that they ought to. In short, one cannot be morally required to be virtuous, if doing so is impossible. Of course this requires much in the way of qualification as 'possibility' comes in many flavours. So to be clear, they need to specify what they mean by 'it is impossible' to be virtuous. Obviously, they do not think it is *logically* impossible to be virtuous – their concern is with empirical evidence and not with showing that virtue ethics contains some kind of conceptual contradiction. Nor are they arguing that it is *physically* impossible to be virtuous. Instead, it is a weaker kind of 'impossibility' that their argument invokes. Their argument must be that the data from psychology shows that no-one *is* in fact robustly or consistently virtuous and from this we can infer that it is *unrealistic* or *unreasonable* to expect people to be. Thus, they are arguing that it is a kind of *psychological* impossibility for humans to be virtuous. As they summarise their case, it "looks to be the case that the available systematic empirical evidence is compatible with virtue being psychologically impossible (or at least wildly improbable), and this suggests that the impossibility of virtue is an empirical possibility that has to be taken seriously."[375]

In this form their argument could have teeth: if it turns out to be true that it is impossible to be virtuous, then virtue ethics as a normative theory is in serious trouble. However, as putting the

[375] *ibid.,* p. 121.

argument in these terms highlights, it is difficult to establish that virtue is psychologically impossible in the sense of being unrealistic and unreasonable to expect from people. To establish this requires careful interpretation of the psychological evidence, and also examination of what kinds of claims virtue ethics makes, to see if the two do in fact intersect. Doris and Stich list a number of ways in which the psychological evidence may be inadequate as support for their argument. They note the following: (i) There may be methodological problems with the research that might undermine the results; (ii) The experimental contexts might be so distant from everyday contexts that the results do not generalise to the "real world" (the world which virtue ethics purports to deal with); (iii) There appear to be very few longitudinal behavioural studies that would help assess character traits over long periods of time (seemingly important if you want to study the robustness, persistence, and influence over time of character traits); and finally, (iv) The experiments may be conceptually irrelevant – character traits operationalized may not correspond to the conceptions of character traits in virtue ethics.[376] Doris and Stich however think that any such criticism would "require evaluating a great deal of psychological research"[377] and that "making a charge stick to one experiment or two, when there are hundreds, if not thousands of relevant studies, is unlikely to effect a satisfying resolution of the controversy."[378] In the following section I will briefly examine the five studies that Doris and Stich cite, and argue that there are a number of serious problems with using the results of all five of these particular studies to establish that virtue is psychologically impossible in the relevant sense. Further, if these five studies are in fact representative of the most relevant, striking, or important findings, and all are found lacking, then it casts serious doubt on their claim that there are "hundreds, if not thousands of relevant studies"[379] that will support the same conclusion if their reasoning is the same in each. The burden of proof would then be on Doris and Stich to point to other studies out of these many hundreds that they claim are relevant, and show why they are not vulnerable to similar objections.

[376] Points (i)–(iv) from *ibid*., pp. 121-122.
[377] *ibid*., p. 123.
[378] *ibid*., p. 123.
[379] *ibid*., p. 123.

6.4.2 Does the evidence show what Doris and Stich claim?

To begin, I examine whether it is true that the figures cited by Doris and Stich in the psychological research are suggestive of psychological impossibility or impracticality with regards to the traits in question. The Mathews and Canon study showed that fewer people helped an injured stranger pick up some dropped books, in the presence of a loud faulty power lawnmower (measured at 87dB), compared to when ambient noise was at normal levels (50dB). In this study, 80 percent of people assisted the injured person at normal noise levels, but only 15 percent helped when the lawn mower was running nearby. In the Darley and Batson study, passers-by who were in a hurry were less likely to help someone in distress than those who were not: 63 percent of people assisted in the no-hurry case, but only 10 per cent stopped to give help if they were in a hurry. In the third study cited, conducted by Isen and Levin, 88 per cent of people who had just found a dime stopped to help a woman who had dropped some papers, compared to only 4 per cent who did not find a dime. The first aspect of these results to note that in none of these studies was the percentage of people who continued to act in 'virtuous'/trait-relevant ways (helping pick up dropped books and so on) in the more trying circumstances is zero.

Doris and Stich's summary of Stanley Milgram's experiments on authority and disobedience states: "Milgram found that subjects would repeatedly 'punish' a screaming 'victim' with realistic (but simulated) electric shocks at the polite request of an experimenter." While this description has good shock value, it omits a number of important details as I shall discuss. However, the important point to note for now is that Doris and Stich omit the fact that in Milgram's original experiment, while 26 out of 40 subjects *did* continue to participate in the experiment till its completion, the remaining 14 refused to continue at some point before the experiments' end. Milgram summarises the results as follows:

> Of the 40 subjects, 5 refused to obey the experimental commands beyond the 300-volt level. Four more subjects administered one further shock, and then refused to go on. Two broke off at the 330-volt level, and 1 each at 345, 360, and 375 volts. Thus a total of 14 subjects defied the experimenter.[380]

And similarly, Doris and Stich's summary of the Haney *et al*. study (commonly known as the 'Stanford Prison Experiment') omits relevant details. They claim that "college students role-playing in a simulated prison rapidly descended to *Lord of the Flies* barbarism."[381] This sparse description omits the fact that there were a range of different reactions from individual participants (both those role-playing guards and prisoners). The following is taken from the results of the Haney *et al* study:

The extremely pathological reactions which emerged in both groups of subjects testify to the power of the social forces operating, but still there were individual differences seen in styles of coping with this novel experience and in degrees of successful adaptation to it. Half of the prisoners did endure the oppressive atmosphere, and not all the guards resorted to hostility. Some guards were tough but fair ("played by the rules"), some went far beyond their roles to engage in creative cruelty and harassment, while a few were passive and rarely instigated any coercive control over the prisoners.[382]

Doris and Stich do not mention these differences in response in this study, but since the situational factors are supposedly controlled for, a reasonable assumption would be that differences in disposition and character account for the observed range of responses. So, while the studies Doris and Stich cite show that in a range of contexts it was possible, through relatively small changes in the situation, to alter the proportion of people displaying helpful behaviour or behaviour demonstrating concern towards strangers, there was still significant variation in individual's responses. Importantly, in all five studies mentioned there were a wide range responses to the various experimental situations, and in all studies there *were* people who did *not* act differently or 'non-virtuously' despite the manipulated situational variables.

---

[380] Stanley Milgram, 'Behavioural study of obedience', pp. 375-376.
[381] John Doris, Stephen Stich, 'As a matter of fact: Empirical perspectives on ethics', p. 118.
[382] C. Haney, C. Banks, and P. Zimbardo, 'Interpersonal dynamics in a simulated prison', p. 81.

Given this range of responses in the five cited studies, it is worth considering how strong Doris and Stich's argument about impossibility is. Consider for example, an analogous argument against utilitarianism where the numbers or proportions of people acting in accord with what utilitarianism dictates, are similar to the numbers or proportions that are not influenced by situational variables to 'act against their character' in the studies cited by Doris and Stich. Suppose that a utilitarian argues that we ought to contribute a relatively small amount of money to foreign aid. By contributing what would be a fairly insignificant amount (for those living in affluent nations), it would be possible to alleviate a great amount of suffering and loss of life at little cost. Now suppose that some empirical research is done to find out how many people actually do contribute to such foreign aid, and it is discovered that only a very small percentage of the population do. Would such a finding show that the utilitarian's theory was seriously flawed? The utilitarian would surely find such an argument underwhelming. The fact that some practice is not widely instantiated is not sufficient for us to infer that it is unrealistic to expect that it *could* become widespread. The fact that systematic discrimination and sexist attitudes towards women were once prevalent should not have caused us to infer that it was *impossible* for men to hold different attitudes. Indeed, for the possibility of moral progress at all there must be a difference between the way the world is, and what a moral theory prescribes. If this gap becomes too small, the moral theory risks being too conservative and not advocating any change at all. If a possibility for change is only considered realistic (and thus not ruled out because we have no obligation to do what is impossible) when there is already a majority of people acting in that way, then it would be hard to find ways of changing things for the better that are not immediately ruled out as psychologically impossible or implausible.

Doris and Stich however, think that such a reply has questionable applicability in the case of virtue ethics. They think that virtue ethics, as it is generally conceived, has a special interest in moral education, and that if virtuous character traits are expected to be rare, then "it is not obvious what

role virtue theory could have in a (generally applicable) programme of moral education."[383] They point out that writing on virtue ethics has traditionally had a distinctive emphasis on moral education, where moral education is "construed as aiming for the development of the good character necessary for a good life" [384] And thus they think that virtue cannot be restricted to "a few extraordinary individuals"[385] if it is to fulfil its purported role. Were it true that the evidence shows that only "a few extraordinary individuals" display virtue, this may be an adequate response, but it is not clear that the evidence *does* show that virtue is to be found only in "a few extraordinary individuals." Their use of this phrase is questionable based on the numbers in the research they cite: 15% in the Mathews and Canon study, 10% in the Darley and Batson study, 4% in the Isen and Levin study and significant numbers in the Milgram (arguably 35%) and Haney *et al*. studies. Proportions in these last two studies are not easily quantified in a simple percentage figure, as there was a range of responses to the experimental conditions, some displaying more or less susceptibility to situational influences. So, whether the figures support Doris and Stich's claim that virtue is too rare, depends on what proportion would be sufficient to play a "generally applicable role in a programme of moral education", but it certainly does not appear that virtue is limited to only "a few extraordinary individuals"[386]. There are however, more significant problems with applying the research Doris and Stich cite to virtue ethics as I discuss in the following.

One such difficulty is that Doris and Stich do not specify which trait or virtue each study purports to show to be ineffectual in ensuring virtuous behaviour in the experimental contexts. The problem with not specifying what trait is relevant to each experiment is that we cannot evaluate whether such a trait is of the kind that virtue ethics actually assumes people do or could have. For example, we could guess that for the first three cited studies, Doris and Stich have in mind a trait we might call

[383] *ibid*., p. 120.
[384] *ibid*., p. 120.
[385] *Ibid.,* p. 120.
[386] Consider for example a small country such as New Zealand with a population of approximately 4.2 million people: of these 15% amounts to 630,000, and 4% to 168,000 – these figures appear too high to imply psychological impossibility.

'helpfulness': the propensity or willingness to help others. Thus, if we imagine our perfectly 'helpful' person, it is possible to think that they might always be helpful to complete strangers, even in demanding situations (say, when they were late to fulfil some other obligation, or when they would be required to assist someone in front of a running 87dB lawnmower). But it is important to note just how demanding this conception is. While each individual case of helping might not be overly demanding, consistently being generous and helpful to everyone at every possible opportunity is *very* demanding. Someone who always helps strangers, at every opportunity, is someone that will have very little or no time for their own projects or lives, as there are almost limitless opportunities to help others. While some people may live up to this standard, and dedicate all their time and effort to helping others (the "Mother Teresa's" of the world), the number of such people will very likely amount to at best, "a few extraordinary individuals". Thus, it appears that this conception is too demanding to realistically be what virtue ethics prescribes for everyone.

A less demanding conception of 'helpfulness' or the 'helpful person' may be a more plausible candidate for the version of 'helpfulness' used in virtue ethics. Consider the person who is always helpful to family members, to those they work with, to those they live with, to friends, and so on. This is a character trait that seems like it could be plausibly robust and consistent: people's commitments to their families, friends, and so on, appear to be of the kind we would expect to be unwavering and reliable regardless of 'insubstantial' changes in the situation. Therefore, it is possible that someone might be a generous or helpful person in this more limited sense, and yet only help strangers on rare occasions or when the situation is relatively dire. Thus, for the first three studies Doris and Stich cite, it is possible that the conception of helpfulness that the studies test for, is too demanding to be the kind that virtue ethics is concerned with.[387] If Doris and Stich were more specific about which traits each study was supposed to inform us about, their argument could be considerably strengthened as they would be able to specify what kind of behaviour we would expect those with the traits in question

[387] I discuss later in more detail what virtue ethics' conception of virtues are, but at this point it suffices to note that the conception Doris and Stich appear to be assuming, is too demanding.

to display, and we could then straightforwardly see if the psychological experiments confirmed or disconfirmed that such a trait was efficacious in producing such 'trait-relevant' behaviour. As it stands though, we are left to guess at what the traits in question are, and the most obvious guess appears to be too demanding to be a trait that virtue ethics assumes.

The character traits in question therefore, need to be carefully specified, so that it can be clearly seen whether the character traits that the psychological experiments tell us about are the same as those virtue ethics is concerned with. Further, it must be clear that the studies do not test only for extreme versions of character traits: one can be generous without giving away everything one owns, or helpful without helping everyone at any possible opportunity. For the first three cited 'helpfulness' cases therefore, it appears that the conception of virtue that Doris and Stich are arguing against does not exist; it is one that is too demanding for virtue ethicists to have claimed *should* or *would* be widespread. While some virtue ethicists may argue for such extreme versions (just as some utilitarians argue we should continue to give aid up to the point just before where we would be sacrificing something of equal moral significance) the failure of such a demanding conception of virtue ethics cannot be generalised to the failure of all or any possible version of virtue ethics. A more limited conception of helpfulness whereby helping is expected towards friends, family, colleagues, neighbours, and so on, is a more reasonable form of 'helpfulness' to expect.[388]

I next consider the Milgram and Haney *et al*. studies which Doris and Stich take to support their argument. The first thing to note is that these studies do not focus on any one particular trait, and as discussed for the first three studies, this is a serious problem for Doris and Stich. Beginning with the Milgram study, it is not obvious what particular trait-relevant behaviour is being shown to not occur. Milgram's research was aimed at exploring obedience to authority. It involved a naïve subject, the experimenter, and a confederate of the experimenter. The test subjects were those who answered a

[388] It may be objected that this sets the bar too low for possessing virtue. But of course, there *are* people that fail at these things, and these are precisely those people that virtue ethics would contrast with the virtuous person, and label as unhelpful, selfish, self-seeking, uncooperative, uncaring, inconsiderate, un-neighbourly and so on.

newspaper advertisement looking for volunteers to take part in a psychology experiment ostensibly set up to study the effects of punishment on learning and memory. Upon arrival, the volunteer and confederate (who the volunteer thought was another willing volunteer) drew slips of paper out of hat to determine who would be the 'teacher' and who would be the 'learner'. The draw was rigged so that the volunteer would always be the 'teacher'. With the volunteer in one room with the experimenter, and the confederate in an adjacent room, the volunteer would read out lists of word pairs, and ask the 'learner' to recite them back. If they responded correctly, the volunteer moved on to the next list, if they made a mistake, the volunteer was required to push a button on a "shock machine" to shock the 'learner', starting with the smallest level of shock, and increasing with each wrong answer (the 'learner' pretended to be shocked – no real shocks were administered other than a 'trial' shock which the volunteer underwent to help convince them of the reality of the study). After a number of increases, the 'learner' would begin to cry out with each shock and bang on the wall. If the shocks continued the 'learner' would then complain of a heart condition, and then finally cease making any noises. Volunteers were told that the electric shocks are "painful but not dangerous"[389] and if they questioned if the 'learner' was in danger they were reassured by the experimenter that "Although the shocks may be painful, there is no permanent tissue damage, so please go on."[390]

As mentioned previously, the result of this study was that 14 of the 40 subjects defied the experimenter and ceased to obey the experimenter's commands at some point before reaching the maximum level of electric shock, and the remaining 26 continued to shock the 'learner' to the maximum level. In the original paper, Milgram provides two examples of remarks, transcribed from the tape recordings, made by subjects who refused to continue:

> [0124] "I think he's trying to communicate, he's knocking…. Well it's not fair to shock the guy… these are terrific volts. I don't think this is very humane…. Oh, I can't go on with this; no, this isn't right. It's a hell of an experiment. The guy is

[389] Stanley Milgram, 'Behavioural study of obedience', p. 378.
[390] *ibid.,* p. 374.

> suffering in there. No, I don't want to go on. This is crazy." [Subject refused to administer more shocks.]
>
> [0123] "He's banging in there. I'm gonna chicken out. I'd like to continue, but I can't do that to a man…. I'm sorry I can't do that to a man. I'll hurt his heart. You take your check…. No really, I couldn't do it." [Subject refused to administer more shocks.]

Presumably Doris and Stich think that the perfectly virtuous person in this situation (according to what virtue ethics might recommend) ought to decline to take part in the study if it were apparent that it would involve acting against their character. Or if it were not apparent before participation begins that the study involves acting in 'non-virtuous' ways, that as soon as it *did* become apparent, they would opt out. Simple participation in the experiment therefore cannot be faulted as being 'non-virtuous' in any sense, as up until the point where the 'learner' starts complaining, volunteers think that the 'learner' was there of their own volition and that shocks were not harmful. Only once the experiment was underway, and it *appeared* that the shocks were more serious than the experimenter was letting on, would we expect the 'virtuous person' to leap into (virtuous) action. This is not to deny that it is surprising and disturbing that 65% of participants did continue to shock the 'learner' at the request of the experimenter once the 'learner' was complaining and banging on the wall. But for the purposes of showing that some participants dispositions did not allow them to continue, the important fact is that the remaining 35% did opt out of the experiment.

In addition to it not showing that virtue is impractical or even statistically rare, the other serious difficulty with using Milgram's research is that again we do not know which character traits are supposed to be shown to be ineffectual in producing virtuous behaviour. The problem is that it appears that there could be a number of possible character traits involved, and further, that these could well be in *conflict*. If this is the case then the Milgram study is poorly suited to showing that character traits do not exist. If there is a choice of what to do, and different choices would be acting in accord with different virtues, then the fact that a majority of people take one option over the other does not show that both virtues do not exist or are never instantiated. Instead it may be that acting in

accord with both virtues is impossible as the required actions are mutually exclusive or that the dilemma is one in which there is no 'correct' answer – even for someone with perfect understanding of virtue and practical wisdom – it may simply be an irresolvable dilemma. In the discussion of the cited study, Milgram himself provides thirteen reasons which he thinks contribute to explaining the high level of obedience of volunteers in his study and of these, at least eight directly reference conflicts of various obligations or character traits. These include the following (numbering from the original paper):

> 2. The experiment is, on the face of it, designed to attain a worthy purpose – advancement of knowledge about learning and memory. [The subjects'] obedience occurs not as an end in itself, but as an instrumental element in a situation that the subject construes as significant and meaningful. He may not be able to see its full significance, but he may properly assume that the experimenter does.
>
> 3. The subject perceives that the victim has voluntarily submitted to the authority system of the experimenter…He has taken the trouble to come to the laboratory presumably to aid the experimental research…Thus he has in some degree incurred an obligation toward the experimenter.
>
> 4. The subject, too, has entered the experiment voluntarily, and perceives himself under obligation to the experimenter. He has made a commitment, and to disrupt the experiment is a repudiation of this initial promise of aid.
>
> 5. Certain features of the procedure strengthen the subject's sense of obligation to the experimenter. For one, he has been paid for coming to the laboratory.
>
> 8. Subjects are assured that the shocks administered to the subject are "painful but not dangerous." Thus they assume that the discomfort caused to the victim is momentary, while the scientific gains resulting from the experiment are enduring.
>
> 9. Through Shock Level 20 [out of 30 levels] the victim continues to provide answers on the signal box. The subject may construe this as a sign that the victim is still willing to 'play the game.'
>
> 10. The subject is placed in a position in which he must respond to the competing demands of two persons: the experimenter and the victim. The conflict must be

> resolved by meeting the demands of one or the other; satisfaction of the victim and the experimenter are mutually exclusive…Thus the subject is forced into a public conflict that does not permit any completely satisfactory solution.
>
> 13. At a more general level, the conflict stems from the opposition of two deeply ingrained behaviour dispositions: first, the disposition not to harm other people, and second, the tendency to obey those whom we perceive to be legitimate authorities.[391]

Thus, a reasonable interpretation of subjects' behaviour, based on Milgram's discussion, is that subjects felt both obligations towards the experimenters, and an obligation to the 'learner' to avoid harming them. On this interpretation, the results of this study therefore would be not that it shows that people do not have robust dispositions of the kind virtue ethics is concerned with, but that they have a multitude of such dispositions and that in 35% of the subjects, the disposition to avoid harming others was strong enough for them to overcome various other dispositions of character. These other dispositions might for example be: helping those to whom one has made a commitment to help (the experimenter), not breaking contracts that one enters (being paid for their time), and listening to the advice of authority figures, who are in a position of authority due to their knowledge of a subject, when they claim something (such as that the shocks are painful but not dangerous). Thus, looking in more detail at the cited study shows that it does not provide good evidence for the fact that virtue does not exist.

In the Haney *et al*. study, male students responded to a newspaper advertisement to take part in a psychological study of prison life. Out of the respondents, the 24 who were "Judged to be the most stable (physically and mentally), most mature, and least involved in antisocial behaviour" were chosen to take part. The experiment involved half of the volunteers (selected randomly) playing the role of guards and the other half playing the role of prisoners in a simulated prison environment constructed in the basement of a psychology building. All the volunteers quickly became absorbed in their roles,

[391] Stanley Milgram, 'Behavioural study of obedience', pp. 377-378. The numbers are from the study, with the non-relevant numbers omitted.

some taking them far more seriously than had been predicted. The 'prisoners' received harsh and humiliating treatment from the 'guards', with a number of the guards appearing to find it gratifying, as Haney *et al.* write:

> It was clear that they enjoyed the simple act of controlling some other person. They were corrupted by the power of their roles and became quite inventive in their techniques of breaking the spirit of the prisoners, making them fell worthless.[392]

The study was scheduled to run for two full weeks, but was terminated after only six days, due to the "unexpectedly intense reactions" produced by the mock-prison situation. Similar criticisms as those that apply to the Milgram study (discussed above), also apply to using this study in Doris and Stich's argument against virtue ethics. As mentioned earlier, there was not a uniform response to the experimental conditions, some of the 'guards' appeared to take to the role more willingly or with more pleasure than others:

> About a third of the guards became tyrannical in their arbitrary use of power... Some of the guards merely did their jobs; in fact, the prisoners said they were tough, but fair, correctional officers. Several were good guards. By this they meant they did small favours, they were friendly; they told the prisoners their names.[393]

Thus, while the Stanford prison study may show that *some* apparently 'ordinary' people who find themselves in the 'wrong' circumstances may easily end up behaving in less than desirable ways, it does not show that they all did. Instead, if anything, it showed there *was* a range of responses to the new roles the prisoners and guards found themselves in. One of the commonly made criticisms of the study is that Haney *et al.* do not adequately explain this variance in response. In addition to this, there are a number of problems with the scientific validity of the Stanford prison experiment, which often lead psychologists to treat its results with caution as I discuss below.

[392] Philip Zimbardo, 'The power and pathology of imprisonment', p. 112.
[393] *Ibid.*

Firstly, it has been suggested that students who volunteer for a study of prison life are more likely to possess dispositions associated with behaving abusively. Thomas Caranahan and Sam McFarland ran a study that used a newspaper advertisement virtually identical to that used in the Stanford Prison Experiment, and an identical advertisement that simply stated it was for a psychological study with no mention of prison life.[394] Those who volunteered for the prison study "scored significantly higher on measures of the abuse-related dispositions of aggressiveness, authoritarianism, Machiavellianism, narcissism, and social dominance and lower on empathy and altruism, two qualities inversely related to aggressive abuse".[395] Caranahan and McFarland conclude that "although implications for the Stanford prison experiment remain a matter of conjecture, an interpretation in terms of person-situation interactionism rather than a strict situationist account is indicated by these findings."[396] So it is possible self-selection resulted in a sample of volunteers that already had stronger dispositions to behaving abusively than the norm.

Secondly, there are difficulties with the level of participation by the experimenters and with their expectations influencing the volunteers' behaviour. One of the experimenters' motivations for the Stanford prison experiment was to confirm their situationist view that seemingly ordinary people could be easily influenced by 'bad situations' to act in less than commendable ways. Thus, participation and instructions given by the researchers had the potential to influence the outcome of the study in the direction of the desired result. The researchers played an active role in the experiment, with Philip Zimbardo – the lead researcher – playing the role of "prison superintendent". The experimenters participated in a number of the events that took place, and although they did not actively direct the guards or prisoners much of the time, they made clear their expectations of what the guards were to do. The following is an excerpt from the instructions given to guards:

[394] Thomas Caranahan, Sam McFarland, 'Revisiting the Stanford prison experiment: Could participant self-selection have led to cruelty?'
[395] *ibid*., p. 603.
[396] *ibid*., p. 603.

> You can create in the Prisoners feelings of boredom, a sense of fear to some degree, you can create a notion of arbitrariness that their life is totally controlled by us, by the system, you, me, and they'll have no privacy… They have no freedom of action they can do nothing, say nothing that we don't permit. We're going to take away their individuality in various ways. In general what all this leads to is a sense of powerlessness. That is, in this situation we'll have all the power and they'll have none.[397]

Thus, it was obvious to volunteers that certain outcomes of their behaviour were expected as part of the experiment. The result of this is that it is not clear how much of the 'bad' behaviour of the prisoners or guards was due to role-playing which they thought was expected of them. The most notorious guard for example, who came to be nicknamed "John Wayne" by the prisoners, modelled his harsh and belligerent behaviour on a prison guard from a movie, even going as far as adopting the Southern accent of the character he was emulating.[398] These points about the possibly biased self-selection of participants, the active participation by those running the experiment, and the question of how much of their behaviour is attributable to role-playing, all weaken both the scientific credibility of the Haney *et al.* study, but also more importantly, they weaken the applicability of study to Doris and Stich's argument. In light of these considerations, it becomes much more difficult to criticise the guards, especially those considered "tough but fair" or "good" guards, as acting non-virtuously.

6.4.3 Differing conceptions of virtue

A final reply to Doris and Stich's argument is that their conception of what a virtue is differs from the conception of virtues used by modern virtue ethicists. Doris and Stich take the following as their understanding of virtues:

> As we understand the tradition, virtues are supposed to be robust traits; if a person has a robust trait, she can be confidently (although perhaps not with absolute

[397] S. Alexander Haslam, Stephen Reicher, 'Beyond Stanford: Questioning a role-based explanation of tyranny', p. 22.
[398] C. Haney, C. Maslach, P. Zimbardo, 'Reflections on the Stanford prison experiment: Genesis, transformations, consequences'.

> certainty) expected to display trait-relevant behaviour across a wide variety of trait relevant situations, even where some or all of these situations are not optimally conducive to such behaviour.[399]

Modern virtue ethicists consider virtues to be what they call "multi-track" dispositions. A multi-track disposition is more complex than a "single-track" disposition which is simply a tendency to act consistently in some way. Having a multi-track disposition involves having certain patterns of choices, desires, attitudes, expectations and sensibilities. It is to have a certain kind of "complex mindset" which takes certain kinds of reasons for action as valid reasons. This is best illustrated by an example: consider the virtue of 'honesty'. To have a multi-track virtue of honesty is not simply to be someone who does not cheat and is honest in their dealings. If one does not cheat and is honest simply because the agent fears getting caught and subsequent consequences rather, then they are not 'honest' in the virtue ethicist's sense. Only if they recognise that 'to do otherwise would be dishonest' is the reason they should not cheat or lie, could their actions contribute to our attributing them with honesty. As Rosalind Hursthouse writes

> An honest person cannot be identified simply as one who, for example, always tells the truth, nor even as one who always tells the truth because it is the truth, for one can have the virtue of honesty without being tactless or indiscreet. The honest person recognises "That would be a lie" as a strong (though perhaps not overriding) reason for not making certain statements in certain circumstances, and gives due, but not overriding, weight to "That would be the truth" as a reason for making them.[400]

If this is the conception of virtue that a virtue ethicist holds, then it is unlikely to be shown to be psychologically impossible by the psychological research that Doris and Stich cite. We cannot attribute a multi-track (or lack of a multi-track) trait, based upon a single action or omission. Thus, it would be reckless to ascribe a lack of a demanding trait, such as charity, in cases like the Mathews and Canon

[399] John Doris, Stephen Stich, 'As a matter of fact: Empirical perspectives on ethics', pp. 118-119.
[400] Rosalind Hursthouse, *On virtue ethics,* 'Virtue ethics', sec. 2. and Rosalind Hursthouse, 'Are Virtues the proper starting point for morality?'

or Isen and Levin studies, when all the agents have done is “exhibited conventional decency” rather than acting in a supererogatory manner in a large percentage of cases.

6.4.4 The success of Doris and Stich’s argument against virtue ethics

Doris and Stich argue that evidence from social psychology can be used to show that virtue ethics is committed to a conception of character traits which are psychologically unrealistic or impossible. They present a number of studies which support what is known as ‘situationism’ in psychology. This is the idea that people’s behaviour can be easily influenced by situations they find themselves in: although we may think dispositions or people’s character plays a large role in how people act, this, according to situationism, is false.

However, as I have argued, the evidence Doris and Stich cite for this conclusion is insufficient to establish that people do not or cannot have robust dispositions of the kind they claim virtue ethics requires. The research they cite is not strong enough in terms of raw numbers, they are not clear enough or specific enough about which virtues in each case matter, and some of the research they cite (especially the Milgram and Haney *et al.* studies) could easily being construed as supporting the opposite conclusion: that the best explanation of why some of those behaved as they did, is because of their comparatively virtuous character or disposition. Doris and Stich also do not consider that virtues may conflict which makes some research, such as the Milgram and Haney *et al.* studies, particularly difficult to draw useful conclusions about robust dispositional traits from. Additionally, there are a number of problems with the conception of virtue ethics that Doris and Stich use. No virtue ethicist claims that people instantiate perfect virtue: the conception that appears to be in use by Doris and Stich is simply too demanding. To have a helpful disposition does not entail one helps everyone one comes across to the exclusion of ever doing anything else. Additionally, their conception also appears to be too simple in comparison to what a number of modern virtue ethicists claim. Virtues

are considered to be complex ‘multi-track’ dispositions, rather than ‘single-track’ tendencies to act in a particular way, and this severely limits the applicability of ‘situationist’ argument.

Thus, I conclude that Doris and Stich’s argument based on the ‘situationist’ findings of moral psychology fails to undermine virtue ethics. Nevertheless, Doris and Stich’s attack on virtue ethics is a valuable contribution in that it forces virtue ethicists to focus on the psychological possibility of what they are suggesting, and hopefully encourages them to engage with and consider how well any of their claims fit with, what moral psychology has to tell us about the kind of dispositions people can and do have.

Chapter 7 Further implications for a framework to assess empirical approaches to ethics

In this chapter I continue the undertaking of chapter 4, and further develop the observations and guidance based on the presented case studies. The discussions in the previous chapter examined arguments put forward by Nichols, Roskies, and Doris and Stich who each attempt to argue that empirical findings have important implications for philosophical theories of ethics or metaethics. As in chapter 4 I analyse where they were successful, unsuccessful, and what we can learn from their approach and results. I summarise these findings at the end of a chapter in a table.

7.1 Are Nichols, Roskies, and Doris and Stitch's arguments sound?

Shaun Nichols' arguments focus on varieties of moral rationalism. In doing so he fails to identify perhaps the most important forms of rationalism in the philosophical literature, but nonetheless makes interesting arguments against two varieties which he labels empirical rationalism and conceptual rationalism. His argument against empirical rationalism faces two problems. Firstly, it is only applicable if motivation internalism is true. Secondly, the evidence used to support the moral/conventional distinction used in the research he cites is neither robust or reliable. For the argument to succeed would require more focused empirical research aimed at establishing what he initially argues the evidence shows. Nichols' argument against conceptual rationalism is also unsuccessful, again due to weaknesses in the research cited, but under further examination also due to the interpretation and application to metaethics.

Adina Roskies' argument focused specifically on motivation internalism rather than more general rationalist theories. She cites evidence that initially shows promise for adjudicating between motivation internalism and motivation externalism. However, a number of difficulties with the evidence prevents Roskies' argument form succeeding. Firstly, while the test used to indicate the

presence of motivation is indicative of some kind of emotional response to stimuli, this does not necessarily indicate moral motivation. Secondly, the scenarios presented to VM patients were all hypothetical, thus gauging actual motivation based on their responses to these imagined scenarios is problematic. Finally, even if these weaknesses in the evidence were able to be addressed through new methods and research, there appears to be significant consensus in the wider psychological literature that VM patients' deficiencies include difficulties with making moral judgments and practical reasoning.

Doris and Stich argue that humans cannot realise virtues of the kind virtue ethics hopes for; that its picture of virtue is not in accordance with our empirical understanding of human nature. However, careful review of the research Doris and Stich cite shows that despite initial appearances, their argument does not strongly support the interpretation that virtue cannot exist. Additionally, they are not explicit about which specific virtue is relevant in each case study, which makes evaluating the relevance of the findings problematic. Indeed, some of the research they cite could even be construed as supporting the opposite conclusions to those they reach. Doris and Stich also do not consider that virtues may conflict. This makes it particularly difficult to draw useful conclusions where the studies include opportunities for multiple virtues to be required or exhibited.

Further problems for their argument stem from the conception of virtue their argument assumes. In many cases it appears much too demanding and would require ignoring all other considerations for someone to realise that virtue. Finally, their conception also appears to be too simple in comparison to what a number of modern virtue ethicists claim. Virtues are considered to be complex 'multi-track' dispositions, rather than 'single-track' tendencies to act in a particular way, and this severely limits the applicability of the 'situationist' argument.

Despite these shortcomings, Doris and Stich's attack on virtue ethics forces virtue ethicists to focus on the psychological possibility of the theories they are suggesting and to engage with, and consider how

well virtue ethics' claims fit, with what moral psychology has to tell us about the kinds of dispositions people can and do have.

7.2 Lessons from Nichols on moral rationalism

Shaun Nichols targets a broad category of positions that fall under the name 'moral rationalism' with arguments based on empirical considerations and research. He presents a range of quotes from different moral rationalists (including Thomas Nagel, Christine Korsgaard, and Michael Smith) who he takes to represent the range of moral rationalist positions. From these quotes he extracts two specific moral rationalist theses that he calls empirical rationalism and conceptual rationalism.

One immediate difficulty with Nichols' argument is that neither of these rationalist theses adequately captures the intended ideas of the philosophers he quotes. Instead, the key feature of the rationalist theories he quotes is their aim to *justify* morality through reason or rationality. This highlights the need to ensure that when claiming to argue against a particular philosophical position held by someone, it is important to characterise those positions accurately if the argument is to successfully apply to their theories. Despite this apparent mis-targeting of his arguments, the discussions of the forms of rationalism – empirical and conceptual – are nonetheless philosophically interesting.

However, Nichols' argument against empirical rationalism, the idea that moral judgments are produced by our rational faculties, has a number of problems. Firstly, it assumes a motivation internalist conception of moral judgment; that motivation is a conceptually necessary feature of moral judgment. Thus, before the impact of potential empirical considerations have a chance to be weighed, there is a prior conceptual issue that must be settled as motivation externalists are unlikely to find the argument persuasive. Secondly, Nichols' argument relies on a particular empirically tested distinction – the moral/conventional distinction. This distinction originated in developmental psychology research using 'school yard' type transgressions with primary school aged subjects. Later research has

pointed out significant weaknesses in the application of the same methodology to psychopaths when a wider range of transgressions involving a victim being harmed are presented to adult subjects.

The idea that Nichols identifies as conceptual rationalism will be more familiar to philosophers as the theory of moral motivation internalism. Identifying it as such is helpful as it makes clear its philosophical context in the literature. There are a number of problems with Nichols' argument against motivation internalism that stem from the content and methodology of the surveys his study used, rather than from the validity of the overall approach. Perhaps the most problematic feature of the surveys is that their subject – motivation internalism – is not addressed directly or explicitly. The survey participants are not told that the characters in the thought experiments they partook in lack moral motivation, instead they are told the character John "just doesn't care if he does wrong things." Additionally, the assumption that John should be imagined as practically rational is not explicitly made. Instead participants are told only that the character is of "normal intelligence" – leaving much room for impugned irrationality of various kinds to be interpreted in his actions and character. A further problem is the source and makeup of the population Nichols surveyed – it is difficult to conclude anything definitive about everyone's view of a concept when only one small slice of a single demographic (a class of undergraduate students at one Western university) is surveyed.

In an attempt to replicate and refine Nichols' results, Caj Strandberg and Fredrik Björklund undertook their own surveys which were revised and expanded to address the issues with Nichols' research. Strandberg and Björklund were explicit in asking whether the characters in their thought experiments were morally motivated by the judgments they made, and attempted to account for whether the characters were considered practically rational by participants by comparing situations where practical rationality would likely be compromised and control situations where they would be considered rational.

Strandberg and Björklund also reduced the likelihood of differing interpretations of the survey questions by keeping them simple, atomic, and explicitly about the question of whether motivation

was present when a moral judgment was made by a rational (or irrational) subject. Their sample was much larger and attempted to control for pre-existing philosophical intuitions by ensuring approximately half of the survey participants had some philosophical background and the other half did not.

While the approach of Strandberg and Björklund's surveys improved on the methodology of Nichols' survey, there are still issues with the overall argument due to their interpretation of the results and conclusions they reach. The survey results showed that in most cases the majority of respondents thought it was possible to be rational and make moral judgments without those judgments being motivating. This evidence showed that a higher proportion of respondents had intuitions that supported moral motivation externalism. Their interpretation of this evidence however was that any kind of majority of responses should be considered as showing motivation externalism to be conceptually true. This leap from a mixed response (with a majority) to a wholesale or decisive victory is a misconstrual of the actual evidence and does not support the strength of the conclusion they reached.

The mixed responses to the survey could indicate either that respondents' understanding of the concepts simply were varied, or it could indicate that some respondents interpreted the scenarios and questions differently from each other. Or, a combination of these two options could further confound the results. While Strandberg and Björklund's surveys did attempt to control for varying interpretations, there is no easy way to determine if this was successful and if it was conceptions of moral judgment and moral motivation that varied or if the results were due to the survey design and its interpretation.

7.3 Lessons from Adina Roskies on moral motivation internalism

Adina Roskies argued that subjects with brain injuries to the ventromedial prefrontal cortex (VM patients) are real-life 'rational amoralists' that provide counter examples to moral motivation

internalism. Unlike Nichols, Roskies explicitly locates her argument as targeting the philosophical 'internalist thesis' of the literature. Roskies indicates how her argument differs from similar previous counter examples to motivation internalism, and addresses why some of the common objections to prior attempts are not applicable. For example, she illustrates why it is more difficult to argue that VM patients only make inverted commas moral judgments than it is for the traditional rational amoralist. Unlike the prototypical amoralist from earlier literature, VM patients do not have any reason to be deceptive, and in almost all cases, it is not disputed that prior to their injuries, they had a normal mastery of moral concepts.

However, despite the theoretical promise, Roskies' argument faces a number of difficulties. The first is that the testing methodology used to measure VM patients' motivation following moral judgments is flawed. Roskies takes Skin Conductance Response (SCR) tests as an indication of the presence of moral motivation. But this interpretation is not backed up by the intended usage of SCRs in the research she cites or the wider theoretical understanding of SCRs within psychology and neuroscience. Secondly, all of the research that Roskies cites involved questioning VM patients about hypothetical moral situations, then attempting to measure their motivation following the judgments they made. This testing therefore would not assess actual motivation to act on a moral judgment made by the VM patient, but instead at best could evaluate whether they imagine they would be motivated in the hypothetical situation. This is an important difference as it means the testing Roskies cites does not adequately operationalize the phenomenon of moral motivation in response to moral judgment. Instead, as Kennett and Fine identify in their response to Roskies' argument, correct operationalization would require a testing procedure where a first person *in situ* moral judgment is made by VM patients and their motivation to act on that judgment be measured.

Roskies accepts these criticisms, but responds that both of these problems are methodological issues due in part to the original purpose of the research, which was not intended to support an argument about motivation internalism. With the right methodology for testing the presence of motivation and

using in situ first person moral judgments instead of hypothetical scenarios, it would be possible to remedy these shortcomings. This is a valid response, and the argument would be worth pursuing further if there were no further issues.

However, in response to the suggestion that methodological issues could be solved, Kennett and Fine also draw attention to a different difficulty with Roskies' use of VM patients as the 'rational amoralist' of the literature. Namely, that VM patients do not display a normal capability for making moral judgments. In contrast to Roskies' claims and the evidence she cites, the moral and social judgments VM patients make as part of their everyday lives are often dysfunctional. Phineas Gage is cited as one example of a VM patient but his case is unreliable. Apart from it being unclear whether he actually counts as a VM patient, his example is illustrative of the tendency to read into or interpret very well-known case studies as results that support your conclusions. Gage's story has been used to support a large number of theories, despite the actual evidence and documentation about his case-history being very limited. More generally it is clear that VM patients do not have the well-functioning ability to make moral judgments that Roskies' argument requires. The established view in the psychological literature is that VM patients have general and social decision-making deficiencies. Roskies' interpretation of their moral judgment making capabilities was taken from selected studies rather than the wider literature, so her interpretation of that evidence is not necessarily wrong, but it is an incomplete picture. Establishing VM patients do have a normal capacity for judging is one of the key premises of the argument, so looking more widely to establish this premise would have been advisable.

7.4 Lessons from Doris and Stich on virtue ethics

Doris and Stich claim that virtue ethics is committed to a characterisation of human psychology which features robust character traits. They argue that 'situationist' research from social psychology shows this view of psychology to be incorrect, that people do not have these kinds of robust character traits

and therefore virtue ethics is based on false assumptions. Their conclusion is that at best virtue ethics requires serious revision to be a viable theory and at worst should be rejected outright.

It may be tempting to respond to this argument that virtue ethics' focus is on the kind of virtues people should aim to manifest rather than its legitimate concern being with the kind of virtues people currently do possess. However, once the form of Doris and Stich's argument it is clarified, it becomes apparent that this objection is inadequate. The form of Doris and Stich's argument is a 'cannot implies ought not' argument, which argues that if virtues of the kind virtue ethics recommend are impossible, it cannot be the case that we ought to embody those virtues. This is a valid argument and one of the most promising ways in which empirical results might plausibly result in normative or metaethical conclusions.

There are however two problems that Doris and Stich's argument faces. Firstly, the conception of virtue which they put forward is a very basic one and not one that is representative of the conceptions of virtue supported by modern virtue ethicists. Instead more sophisticated 'multi-track' dispositions are put forward as a more nuanced view of the character traits people have. These multi-track dispositions are much better suited to describing both what people are in fact like and accommodating the empirical evidence Doris and Stich cite.

Secondly, even if virtue ethics did use the basic conception of virtue and character traits that Doris and Stich suggest, it is questionable that the evidence they cite in support of such traits not existing actually shows that virtue is impossible. Their interpretation of the evidence is optimistic in how strongly it supports their argument. Doris and Stich claim that the research shows that no-one is virtuous, but in fact the evidence shows the opposite; that some people are comparatively more virtuous than others. It may be somewhat disappointing that the proportions of people displaying virtue were lower than one might imagine in much of the evidence that is cited, but nonetheless it does not show that character traits or virtue cannot exist.

This result however, is not itself without merit from the point of view of the importance of empirical research for ethical theory. Doris and Stich's argument is a valuable contribution that highlights that environmental effects on behaviour can be very powerful, and that in many cases virtue or character traits are not as influential as our folk conceptions of them might lead us to believe. Their argument pushes virtue ethics towards further developing a more realistic multi-track account of character that includes fewer supererogatory characteristics, and the resulting virtue ethics is a more robust theory for it.

7.5 Additions to a framework for assessing empirical approaches to morality

Drawing from the above considerations from each of the discussions of Shaun Nichols, Adina Roskies, and John Doris and Stephen Stich, the following ideas are provided as additional guidelines for assessing empirical and evolutionary approaches to morality.

Lesson / guideline	Case study / source
Identify philosophical theses carefully and be aware of overly general theories or terms. Often talking about philosophical positions using broad terms such as 'rationalism' may be too non-specific to usefully characterise a position.	§6.2.1 'Are Nichols' rationalisms positions held by philosophers' Nichols argues against an idea he identifies as 'moral rationalism' but cites a range of divergent sources talking about similarly divergent ideas.
Asses what philosophical assumptions your argument makes. The conclusion or importance of the argument may be significantly weakened if it is dependent on philosophical theories or premises being true that are themselves controversial or that your argument has not yet established.	§6.2.2.1 'What is wrong with psychopaths?' In arguing against empirical rationalism, Nichols assumes that motivation internalism is true. If this assumption turned out to be false, he would not be able to attribute an undisturbed capacity for moral judgment making to psychopaths and his argument would not succeed.

Evaluate carefully that the methodology of the empirical research is robust and appropriate for the philosophical argument it is used in.	§6.2.2.2 'The moral/conventional distinction' The research that introduced the moral/conventional distinction was undertaken in developmental psychology. Because of this context, schoolyard transgressions were used that may not have had the moral importance required to distinguish them from the conventional in the later research undertaken on psychopaths.
Where possible use standard names and terminology for philosophical positions to make them more easily recognisable and to better contextualise them within the literature.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Nichols' conceptual rationalism is a form of moral motivation internalism but this is not initially clear from his description of the position. In contrast, Roskies' argument is immediately clearer than Nichols' due to her identifying that her argument targets moral motivation internalism.
Surveys need to be targeted to the specific desired philosophical outcomes when crafting survey questions. Concepts that you wish to distinguish between need to be carefully addressed and targeted in the surveys/questions.	§6.2.3 'Conceptual rationalism and moral motivation internalism' The questions in Nichols' surveys left a lot of room for interpretation by participants and in some cases did not even mention the target concepts explicitly, resulting in it being unclear whether participants had the concepts Nichols was investigating in mind when responding.
Be wary of drawing conclusions from limited or uniform samples – especially if you're trying to conclude something that is supposed to apply to a whole population or the concept used by everyone.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Nichols draws conclusions that are supposed to apply to all usages of a concept from a survey of a single sample of undergraduate students at one western university.
Surveys that show mixed results should not be interpreted as conclusive support for an argument without analysis to support that conclusion. The analysis should acknowledge and explain the distribution of responses as part of the argument supporting the conclusions reached.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Surveys showing a 42/58% split of responses such as responses to Strandberg and Björklund's 'psychopath' vignette, should not be interpreted as decisive or conclusive evidence. Such a result seems more likely to indicate either a lack of shared conception or a question that can be interpreted in multiple ways.

Lack of consensus may not be due to survey methodology or interpretation; an understanding of a concept not being uniformly shared or applied by differing individuals or groups may be a reality philosophy has to live with. Arguments should be open to the result that there is diversity in how a concept is used or understood.	§6.2.3 'Conceptual rationalism and moral motivation internalism' The conclusion of Strandberg and Björklund's surveys is that there is considerable variability of concepts of moral judgment and the modality of the link between moral judgment and moral motivation within philosophically untrained undergraduates. This may imply that there is no definitive or single answer to the question of whether moral motivation is a necessary feature of moral judgement.
Make clear how the argument being presented differs from the previous attempts or similar arguments made in the literature and why the new approach avoids previous objections or difficulties.	§6.3 'Adina Roskies and motivation internalism' Roskies indicates how her argument differs from similar previous counter examples to motivation internalism, and addresses why some of the common objections to prior attempts are not applicable. For example, it is more difficult to argue that VM patients only make inverted commas moral judgments. Unlike the prototypical amoralist from the previous literature, VM patients do not have any reason to be deceptive, and in most cases, it is not disputed that prior to their injuries, they had a normal mastery of moral concepts.
Ensure that the research methodology correctly and robustly assesses the concepts you are trying to test.	§6.3 'Adina Roskies and motivation internalism' The testing methodology Roskies used to assess VM patients' motivation following moral judgments was flawed. Roskies takes Skin Conductance Response (SCR) tests as an indication of the presence of moral motivation. But this interpretation is not backed up by the intended usage of SCRs in the research she cites or the wider theoretical understanding of SCRs within psychology and neuroscience.

Ensure that the research cited is testing the same precise concepts that the argument in philosophy requires. Often empirical research that is suggestive of philosophical conclusions requires revision and explicit targeting to be applicable to the philosophical arguments it prompts.	§6.3 'Adina Roskies and motivation internalism' Roskies cites research that used hypothetical moral scenarios. These are not sufficient for assessing motivation in response to a moral judgment. Instead the research would require first person in situ moral judgments be tested for subsequent moral motivation.
Just because particular examples of research are unsuccessful in supporting an argument does not mean the approach is flawed if the problems with the research could be remedied or generated in more reliable ways.	§6.3 'Adina Roskies and motivation internalism' Roskies' evidence is shown not to support the argument, but the argument is still potentially sound if the evidence can be found elsewhere or further research that addresses the methodological issues is undertaken.
Ensure that your interpretation and understanding of philosophical or conceptual distinctions or assumptions is representative of the wider theoretical understanding within psychology or the relevant field of research. If the interpretation you adopt is not representative of the wider literature, this difference in understanding should be adequately justified.	§6.3 'Adina Roskies and motivation internalism' Roskies interpretation of the deficits of VM patients suited her argument but is not representative of the general understanding of VM patients within neuropsychology which understands VM patients as having serious deficits in moral judgment making capabilities within the context of their own lives.
Be wary of very well-known examples or case-studies that have more mythology than substance to them and have been interpreted to support conclusions or popular ideas that the evidence does not actually support.	§6.3 'Adina Roskies and motivation internalism' Phineas Gage is well known but the actual documentation of his case history is limited, and it is unclear if he fits the typical VM patient profile.
It's important that the structure of the argument is clear to avoid it being misinterpreted as something obviously unsound.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' The form of Doris and Stich's argument is a 'cannot implies ought not' argument, which tries to show that virtues of the kind virtue ethics recommend are impossible therefore it cannot be the case that we ought to embody those virtues. If this structure is not made explicit it is easy to misinterpret their argument as being a misguided interest in how things are instead of how they ought to be.

Concepts that empirical research shows to be inaccurate need to be the same concept as used by the philosophical theory to make the argument stick.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' Doris and Stich's argument targets an overly simple conception of virtue ethics which is not held by any actual virtue ethicists.
Arguments that overall are not successful can still contain important but more modest conclusions that are important and move debates along.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' Doris and Stich's argument may not result in the rejection of virtue ethics, but they do force it to adopt a more robust empirically based conception of virtue, and this is still a notable conclusion.

Chapter 8 Conclusion

8.1 Empirical approaches to moral philosophy

In this thesis I have answered the question '*What are the implications of our growing understanding of the science of morality for ethics?*' I have argued that there is no *a priori* principle that allows us to either rule out empirical research having implications for moral philosophy or to discern whether there are important implications for moral philosophy.

I have examined a number of case studies which attempt to derive important implications for moral philosophy from empirical research into morality. I have examined a representative rather than comprehensive sample of attempts, as the number of potential case studies far exceeds what can be covered in a single thesis. The case studies I have looked at include arguments about the evolutionary origins of morality made by E.O Wilson, Richard Joyce, and Sharon Street. I tentatively concluded that Joyce and Streets' arguments were successful in establishing that an entirely mind independent moral reality is unlikely, and that their arguments should constrain the possible metaphysics of morality to more naturalistic and contingent varieties.

I have also examined a range of research conducted on moral psychology into how moral judgments are made by researchers such as Joshua Greene, Marc Hauser, Jonathan Haidt, and Shaun Nichols. I have discussed Shaun Nichols' arguments about the implications such models have for a number of positions he terms 'moral rationalism'. I argue that while there is nothing wrong in principle with Nichols' approach, ultimately his arguments are unable to establish that the various forms of rationalism are incorrect due to a stalemate of sorts because of the indeterminate nature of some of the concepts involved. I also looked at Adina Roskies' attempt to use empirical research to show that motivational internalism is false. I argue that again the concepts employed are not sufficiently determinate to establish her conclusion and that empirical evidence she presents is not as robust as it first appears.

Finally, I looked at attempts by John Doris and Stephen Stich to show that research into 'situationist' social psychology can be used to undermine Virtue Ethics by showing that the concept of Virtue that Virtue Ethics attributes to people is not possible. While I conclude that their argument is unsuccessful in showing that Virtue Ethics is fatally flawed, they are successful in pushing Virtue Ethics towards more realistic and detailed pictures of what human virtue is like. Thus, their argument is still a highly valuable contribution.

Thus, a number of the case studies I have examined have been successful in limited but still important ways. The overall trend is that an empirical approach often precipitates refinement and adjustment of philosophical theories to accommodate the additional scrutiny on the philosophical and factual assumptions moral philosophy makes. My overall conclusion is positive, that while the relation between ethics and empirical approaches to ethics is indeed an "alluring swamp in which any number of scholars have floundered"[401] there is nevertheless much to be gained by interdisciplinary work of this nature and integrating the insights of empirical research into morality with moral philosophy.

8.2 Framework for assessing empirical approaches to ethics

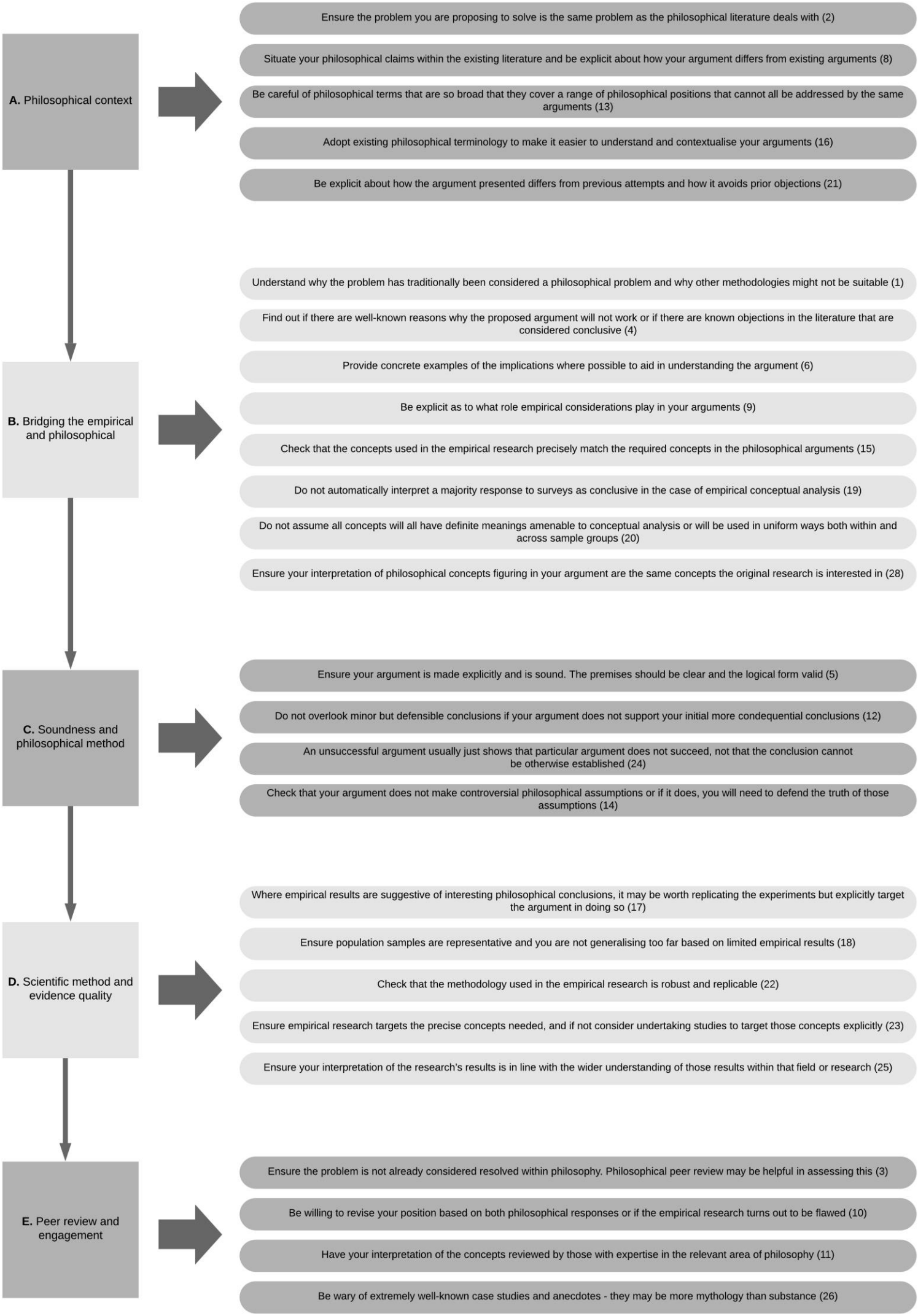
In addition to the above conclusions, I have assessed what we can learn in general from each of case studies' attempts to draw conclusions for ethics from empirical research. I presented these lessons in the tables chapter 4 and chapter 7. As noted in §1.1 the intended audience of this guidance is non-philosophers working on interdisciplinary research involving philosophy, philosophers supplementing philosophy's methodology with the methods and tools of the sciences, or simply non-philosophers who discover their research appears to have philosophical conclusions. While a comprehensive framework based on this approach is too large for a single thesis, the below at least provides a beginning and a direction in which this could be further developed. The guidelines provided in

[401] Philip Kitcher, 'Biology and Ethics', p. 163.

chapters 4 and 7 have been categorised into the following themes based upon the general type of guidance provided:

A. *Philosophical context* – considerations about the existing literature, terminology, arguments, and established philosophical knowledge.
B. *Bridging the empirical and philosophical* – considerations about the forms of arguments used to try to draw conclusions in moral philosophy based on empirical considerations. This includes the 'is-ought gap' and existing philosophical arguments about deriving normative conclusions from the way the world is.
C. *Soundness and philosophical method* – general considerations around the philosophical method and ensuring arguments are sufficiently well developed and are valid and true.
D. *Scientific method and evidence quality* – factors around the quality and applicability of the empirical research used.
E. *Peer review and engagement* – considerations around feedback and the need to engage fully with the philosophical literature and community.

Using these categories, that are intended to be followed loosely in the order presented in the diagram below, a basic pathway is provided that can be used to assist in raising the quality of attempts to draw conclusions for philosophical ethics based on empirical considerations.


A. Philosophical context
Ensure the problem you are proposing to solve is the same problem as the philosophical literature deals with (2)
Situate your philosophical claims within the existing literature and be explicit about how your argument differs from existing arguments (8)
Be careful of philosophical terms that are so broad that they cover a range of philosophical positions that cannot all be addressed by the same arguments (13)
Adopt existing philosophical terminology to make it easier to understand and contextualise your arguments (16)
Be explicit about how the argument presented differs from previous attempts and how it avoids prior objections (21)
B. Bridging the empirical and philosophical
Understand why the problem has traditionally been considered a philosophical problem and why other methodologies might not be suitable (1)
Find out if there are well-known reasons why the proposed argument will not work or if there are known objections in the literature that are considered conclusive (4)
Provide concrete examples of the implications where possible to aid in understanding the argument (6)
Be explicit as to what role empirical considerations play in your arguments (9)
Check that the concepts used in the empirical research precisely match the required concepts in the philosophical arguments (15)
Do not automatically interpret a majority response to surveys as conclusive in the case of empirical conceptual analysis (19)
Do not assume all concepts will all have definite meanings amenable to conceptual analysis or will be used in uniform ways both within and across sample groups (20)
Ensure your interpretation of philosophical concepts figuring in your argument are the same concepts the original research is interested in (28)
C. Soundness and philosophical method
Ensure your argument is made explicitly and is sound. The premises should be clear and the logical form valid (5)
Do not overlook minor but defensible conclusions if your argument does not support your initial more condequential conclusions (12)
An unsuccessful argument usually just shows that particular argument does not succeed, not that the conclusion cannot be otherwise established (24)
Check that your argument does not make controversial philosophical assumptions or if it does, you will need to defend the truth of those assumptions (14)
D. Scientific method and evidence quality
Where empirical results are suggestive of interesting philosophical conclusions, it may be worth replicating the experiments but explicitly target the argument in doing so (17)
Ensure population samples are representative and you are not generalising too far based on limited empirical results (18)
Check that the methodology used in the empirical research is robust and replicable (22)
Ensure empirical research targets the precise concepts needed, and if not consider undertaking studies to target those concepts explicitly (23)
Ensure your interpretation of the research's results is in line with the wider understanding of those results within that field or research (25)
E. Peer review and engagement
Ensure the problem is not already considered resolved within philosophy. Philosophical peer review may be helpful in assessing this (3)
Be willing to revise your position based on both philosophical responses or if the empirical research turns out to be flawed (10)
Have your interpretation of the concepts reviewed by those with expertise in the relevant area of philosophy (11)
Be wary of extremely well-known case studies and anecdotes - they may be more mythology than substance (26)

References

Aiello, Leslie C., and R. I. M. Dunbar. 'Neocortex Size, Group Size, and the Evolution of Language'. *Current Anthropology* 34, no. 2 (1993): 184–93.

Alexander, Richard D. *The Biology of Moral Systems*. Hawthorne, New York: Aldine de Gruyter, 1987.

Allman, J. *Evolving Brains*. Henry Holt and Company, 2000.

Altham, J. E. J., and Ross Harrison. *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge University Press, 1995.

Axelrod, Robert, and William D. Hamilton. 'The Evolution of Cooperation'. *Science* 211 (1981): 1390–96.

Baier, Kurt. 'Why Should We Be Moral?' In *Morality and Rational Self-Interest*, edited by David P. Gauthier. Englewood Cliffs: Prentice Hall, 1970.

Björnsson, Gunnar, Caj Strandberg, Ragnar Francén Olinder, John Eriksson, and Fredrik Björklund. *Motivational Internalism: Contemporary Debates*, 2015.

Blackburn, S. *Spreading the Word: Groundings in the Philosophy of Language*. Clarendon Press, 1984.

Blackburn, Simon. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Clarendon Press, 1998.

Blair, R. J. R. 'A Cognitive Developmental Approach to Morality: Investigating the Psychopath'. *Cognition* 57 (1995): 1–29.

Blair, Robert, Lawrence Jones, F. Clark, and M. Smith. 'Is the Psychopath Morally Insane?' *Personality and Individual Differences* 19 (1995): 741–52.

Bloomfield, Paul. 'Review: The Evolution of Morality'. *Mind* 116, no. 461 (1 January 2007): 176–80. https://doi.org/10.1093/mind/fzm176.

Boniolo, Giovanni, and Gabriele De Anna. *Evolutionary Ethics and Contemporary Biology*. Cambridge; New York: Cambridge University Press, 2006.

Boucsein, Wolfram. *Electrodermal Activity*. 2nd ed. New York: Springer, 2012.

Brian, Medlin. 'Ultimate Principles and Ethical Egoism'. In *Morality and Rational Self-Interest*, edited by David P. Gauthier. Englewood Cliffs: Prentice Hall, 1970.

Brink, Gijsbert van den, Jeroen de Ridder, and René van Woudenberg. 'The Epistemic Status of Evolutionary Theory'. *Theology and Science* 15, no. 4 (2017): 454–72. https://doi.org/10.1080/14746700.2017.1369759.

Browne, Derek. ‘A Science of Morality?, Critical Notice of Marc D. Hauser, Moral Minds (2006)’. *The Rutherford Journal* 2 (2007). http://rutherfordjournal.org/article020107.html.

Cahan, D. *From Natural Philosophy to the Sciences: Writing the History of Nineteenth-Century Science*. University of Chicago Press, 2003.

Carlson, Neil R. *Physiology of Behavior*. Pearson Higher Ed, 2012.

Carnahan, Thomas, and Sam McFarland. ‘Revisiting the Stanford Prison Experiment: Could Participant Self-Selection Have Led to the Cruelty?’ *Personality and Social Psychology Bulletin* 33 (2007): 603–14.

Chalmers, David J. ‘Why Isn’t There More Progress in Philosophy?’ *Philosophy* 90, no. 1 (2015): 3–31. https://doi.org/10.1017/S0031819114000436.

Chomsky, Noam. ‘The Mysteries of Nature: How Deeply Hidden?’ *The Journal of Philosophy* 106, no. 4 (2009): 167–200.

Clarke-Doane, Justin. ‘Morality and Mathematics: The Evolutionary Challenge’. *Ethics* 122, no. 2 (2012): 313–40.

Cohen, G. A. ‘Paradoxes of Conviction’. In *If You’re an Egalitarian, How Come You’re So Rich?* Harvard University Press, 2009.

Copp, David. ‘Darwinian Skepticism about Moral Realism’. *Philosophical Issues* 18, no. 1 (1 September 2008): 186–206. https://doi.org/10.1111/j.1533-6077.2008.00144.x.

———. ‘Moral Naturalism and Three Grades of Normativity’. In *Normativity and Naturalism*, edited by Peter Schaber, 7–45. Frankfurt: Ontos-Verlag, 2004.

———. *The Oxford Handbook of Ethical Theory*. New York: Oxford University Press, 2006.

Cosmides, Leda, and John Tooby. ‘Evolutionary Psychology and the Generation of Culture, Part II: Case Study: A Computational Theory of Social Exchange’. *Ethology and Sociobiology* 10 (1989): 51–97.

Curry, Oliver. ‘Who’s Afraid of the Naturalistic Fallacy?’ *Evolutionary Psychology* 4 (2006): 234–47.

Damasio, Antonio R., Daniel Tranel, and Hanna Damasio. ‘Individuals with Sociopathic Behavior Caused by Frontal Damage Fail to Respond Autonomically to Social Stimuli’. *Behavioural Brain Research* 41, no. 2 (1990): 81–94.

Darley, J. M., and Batson C. D. ‘From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior’. *Journal of Personality and Social Psychology* 27, no. 1 (1973): 100–108.

D'Arms, Justin. 'When Evolutionary Game Theory Explains Morality, What Does It Explain?' *Journal of Consciousness Studies* 7, no. 1–2 (2000): 296–99.

Darwall, Stephen, Alan Gibbard, and Peter Railton. 'Toward Fin de Siecle Ethics: Some Trends'. *The Philosophical Review* 101, no. 1 (1992).

Darwall, Stephen, Allan Gibbard, and Peter Railton. *Moral Discourse and Practice: Some Philisophical Approaches*. New York: Oxford University Press, 1997.

Dawkins, Richard. 'Burying the Vehicle'. *Behavioural and Brain Sciences* 17, no. 4 (1994): 616–17.

———. *The Selfish Gene*. Oxford; New York: Oxford University Press, 2006.

Dennett, Daniel C. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster, 1995.

———. 'E Pluribus Unum?' *Behavioural and Brain Sciences* 17, no. 4 (1994): 617–18.

Deutsch, Max. *The Myth of the Intuitive: Experimental Philosophy and Philosophical Method*. MIT Press, 2015.

Doris, John. *Lack of Character: Personality and Moral Behaviour*. Cambridge, U.K.; New York: Cambridge University Press, 2002.

Doris, John, and Stephen Stich. 'As a Matter of Fact: Empirical Perspectives on Ethics'. In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith. Oxford: Oxford University Press, 2005.

———. 'Moral Psychology: Empirical Approaches'. In *The Stanford Encyclopedia of Philosophy (Fall 2014 Edition)*, edited by Edward N. Zalta, Spring 2014. https://plato.stanford.edu/archives/spr2014/entries/moral-psych-emp/.

Dreier, James, and Rosalind Hursthouse, eds. 'Are Virtues the Proper Starting Point for Morality?' In *Contemporary Debates in Moral Theory*. John Wiley & Sons, 2009.

Dugatkin, Lee Alan. 'The Evolution of Cooperation: Four Paths to the Evolution and Maintenance of Cooperative Behaviour'. *Bioscience* 47, no. 6 (1997): 355–62.

Dunbar, R. I. M. 'Coevolution of Neocortical Size, Group Size and Language in Humans'. *Behavioral and Brain Sciences* 16, no. 4 (1993): 681–94. https://doi.org/10.1017/S0140525X00032325.

Eggers, Daniel. 'Unconditional Motivational Internalism and Hume's Lesson'. In *Motivational Internalism: Contemporary Debates*, edited by Gunnar Björnsson, Caj Strandberg, Ragnar Francén Olinder, John Eriksson, and Fredrik Björklund, 2015.

Enoch, David. 'The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 148, no. 3 (2010): 413–38.

Fellows, L. K. 'Deciding How to Decide: Ventromedial Frontal Lobe Damage Affects Information Acquisition in Multi-Attribute Decision Making'. *Brain* 129, no. Pt 4 (2006): 944–52. https://doi.org/10.1093/brain/awl017.

Fellows, L. K., and M. J. Farah. 'Dissociable Elements of Human Foresight: A Role for the Ventromedial Frontal Lobes in Framing the Future, but Not in Discounting Future Rewards'. *Neuropsychologia* 43, no. 8 (2005): 1214–21. https://doi.org/10.1016/j.neuropsychologia.2004.07.018.

Finlay, Stephen. 'Errors Upon Errors: A Reply to Joyce'. *Australasian Journal of Philosophy* 89, no. 3 (1 September 2011): 535–47. https://doi.org/10.1080/00048402.2010.510531.

———. 'The Error in the Error Theory'. *Australasian Journal of Philosophy* 86, no. 3 (1 September 2008): 347–69. https://doi.org/10.1080/00048400802001921.

FitzPatrick, William. 'Morality and Evolutionary Biology'. In *The Stanford Encyclopedia of Philosophy (Fall 2014)*, edited by Edward N. Zalta, n.d. https://plato.stanford.edu/archives/fall2014/entries/morality-biology/.

FitzPatrick, William J. 'Debunking Evolutionary Debunking of Ethical Realism'. *Philosophical Studies* 172, no. 4 (1 April 2015): 883–904. https://doi.org/10.1007/s11098-014-0295-y.

Flack, Jessica C., and Frans B.M. de Waal. '"Any Animal Whatever": Darwinian Building Blocks of Morality in Monkeys and Apes'. *Journal of Consciousness Studies* 7, no. 1–2 (2000): 1–29.

Francén, Ragnar. 'Moral Motivation Pluralism'. *The Journal of Ethics* 14, no. 2 (2010): 117–48.

Fraser, Benjamin James. 'Evolutionary Debunking Arguments and the Reliability of Moral Cognition'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 168, no. 2 (2014): 457–73.

Gauthier, David P. 'Morality and Advantage'. In *Morality and Rational Self-Interest*, edited by David P. Gauthier. Englewood Cliffs: Prentice Hall, 1970.

———. *Morality and Rational Self-Interest*. Englewood Cliffs: Prentice Hall, 1970.

Gibbard, Allan. *Thinking How to Live*. Cambridge; Mass.: Harvard University Press, 2003.

———. *Wise Choices, Apt Feelings: A Theory of Normative Judgement*. Cambridge; Mass.: Harvard University Press, 1990.

Gigerenzer, G., P. M. Todd, and ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999.

Gill, Michael B. 'Indeterminacy and Variability in Meta-Ethics'. *Philosophical Studies* 145, no. 2 (2009): 215–34.

Goldman, Alvin I. 'Ethics and Cognitive Science'. *Ethics* 103, no. 2 (1993): 337–60.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 'An FMRI Investigation of Emotional Engagement in Moral Judgment'. *Science* 293, 14 September (2001): 2105–8.

Haidt, Jonathan. 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement'. *Psychological Review* 108 (2001): 814–34.

———. 'The New Synthesis in Moral Psychology'. *Science* 316 (2007): 998–1002.

Haidt, Jonathan, Fredrik Bjorklund, and Scott Murphy. 'Moral Dumbfounding: When Intuition Finds No Reason'. *Unpublished Manuscript, University of Virginia*, 2000, 191–221.

Haidt, Jonathan, and C. Joseph. 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues'. *Daedalus* 123, no. 4 (2004): 55–66.

Hamilton, W. D. 'The Genetical Evolution of Social Behaviour'. *Journal of Theoretical Biology* 7 (1964): 1–52.

Haney, Craig, Curtis Banks, and Philip Zimbardo. 'Interpersonal Dynamics in a Simulated Prison'. *International Journal of Criminology and Penology* 1 (1973): 69–97.

Hare, R. M. 'Universalisability'. *Proceedings of the Aristotelian Society* 55 (1954): 295–312.

Harman, Gilbert. 'Skepticism about Character Traits'. *The Journal of Ethics*, 2008.

———. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press, 1977.

Haslam, S. Alexander, and Stephen Reicher. 'Beyond Stanford: Questioning a Role-Based Explanation of Tyranny'. *Dialogue* 18 (2003): 22–25.

———. 'Rethinking the Social Psychology of Tyranny: The BBC Prison Study'. *British Journal of Social Psychology* 45 (2006): 1-40.

Hauser, Marc D. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Ecco, 2006.

Hearnshaw, John. ‘Auguste Comte’s Blunder: An Account of the First Century of Stellar Spectroscopy and How It Took One Hundred Years to Prove That Comte Was Wrong!’ *Journal of Astronomical History and Heritage* 13 (2010): 90–104.

Higgins, Andrew, and Alexis Dyschkant. ‘Interdisciplinary Collaboration in Philosophy’. *Metaphilosophy* 45, no. 3 (1 July 2014): 372–98. https://doi.org/10.1111/meta.12091.

Hobbes, Thomas. ‘The Natural Condition of Mankind and the Laws of Nature’. In *Morality and Rational Self-Interest*, edited by David P. Gauthier. Englewood Cliffs: Prentice Hall, 1970.

Horgan, Terry, and Mark Timmons. *Metaethics after Moore*. Oxford; New York: Clarendon Press, 2006.

Hughes, A. L. *Evolution and Human Kinship*. Oxford University Press, 1988.

Hume, D. *A Treatise of Human Nature*. Dover Publications, 2003.

Hume, David. *An Enquiry Concerning the Principles of Morals*. Edited by Jerome B. Schneewind. Hackett Publishing Company, 1983.

Hursthouse, Rosalind. *On Virtue Ethics*. 1999. Oxford; New York: Oxford University Press, 1999.

———. ‘Virtue Ethics’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, 2008. http://plato.stanford.edu/archives/fall2008/entries/ethics-virtue.

Ingen, John Van. *Why Be Moral?: The Egoistic Challenge*. Peter Lang, 1994.

Isen, Alice M., and Paula F. Levin. ‘Effect of Feeling Good on Helping: Cookies and Kindness’. *Journal of Personality and Social Psychology* 21, no. 3 (1972): 384–88.

Jackson, F., and M. Smith. *The Oxford Handbook of Contemporary Philosophy*. OUP Oxford, 2007.

Johnson, Oliver. *The Moral Life*. London: George Allen and Unwin Ltd, 1969.

Joyce, Richard. ‘Enough with the Errors! A Final Reply to Finlay’. *Unpublished*, 2012. http://personal.victoria.ac.nz/richard_joyce/acrobat/joyce_2012_enough.with.the.errors.pdf.

———. ‘Error Theory’. In *International Encyclopedia of Ethics*, edited by Hugh LaFollette. John Wiley and Sons, 2013.

———. ‘Evolution and Ethics’. In *The Blackwell Guide to Ethical Theory*, edited by Hugh LaFollette and Ingmar Persson, 2nd ed. Blackwell Philosophy Guides, n.d.

———. ‘Irrealism and the Genealogy of Morals’. *Ratio* 26, no. 4 (2013): 351–72.

———. ‘Metaethical Pluralism: How Both Moral Naturalism and Moral Skepticism May Be Permissible Positions’, n.d.

———. ‘Metaethics and the Empirical Sciences’. *Philosophical Explorations* 9, no. 1 (2006): 133 – 148.

———. ‘Moral Anti-Realism’. edited by Edward N. Zalta, 2007. http://plato.stanford.edu/archives/fall2007/entries/moral-anti-realism/.

———. ‘The Error In “The Error In The Error Theory”’. *Australasian Journal of Philosophy* 89, no. 3 (1 September 2011): 519–34. https://doi.org/10.1080/00048402.2010.484465.

———. *The Evolution of Morality*. MIT Press, 2006.

———. *The Myth of Morality*. Cambridge University Press, 2001.

———. ‘What Neuroscience Can (and Cannot) Contribute to Metaethics’. *Moral Psychology* 3 (2008): 371–94.

Joyce, Richard, and Simon Kirchin, eds. *A World without Values*. Springer, 2010.

Kagan, Jerome. ‘Human Morality Is Distinctive’. *Journal of Consciousness Studies* 7, no. 1–2 (2000): 46–48.

Kahane, Guy. ‘Evolutionary Debunking Arguments’. *Noûs* 45, no. 1 (1 March 2011): 103–25. https://doi.org/10.1111/j.1468-0068.2010.00770.x.

Katz, Leonard D. ‘Toward Good and Evil: Evolutionary Approaches to Aspects of Human Morality’. *Journal of Consciousness Studies* 7, no. 1–2 (2000): ix–xvi.

Kelly, Daniel, Stephen Stich, Kevin J. Haley, Serena J. Eng, and Daniel M. T. Fessler. ‘Harm, Affect, and the Moral/Conventional Distinction’. *Mind & Language* 22, no. 2 (2007): 117–31.

Kennett, Jeanette, and Cordelia Fine. ‘Could There Be an Empirical Test for Internalism?’ In *Moral Psychology The Neuroscience of Morality : Emotion, Disease, and Development*, edited by Walter Sinnott-Armstrong, Vol. 3. Cambridge, MA: MIT Press, 2008.

———. ‘Internalism and the Evidence from Psychopaths and “Acquired Sociopaths”’. In *Moral Psychology The Neuroscience of Morality : Emotion, Disease, and Development*, edited by Walter Sinnott-Armstrong, 3:173–90. Cambridge, MA: MIT Press, 2008.

Kihlstrom, John F. ‘Social Neuroscience: The Footprints of Phineas Gage’. *Social Cognition* 28, no. 6 (2010): 757–83. https://doi.org/10.1521/soco.2010.28.6.757.

Kitcher, P. *The Ethical Project*. Harvard University Press, 2011.

Kitcher, Philip. ‘Biology and Ethics’. In *The Oxford Handbook of Ethical Theory*, edited by David Copp. New York: Oxford University Press, 2006.

———. ‘Four Ways of “Biologicizing” Ethics’. In *Conceptual Issues in Evolutionary Biology: Third Edition*, edited by Elliot Sober. Cambridge, Mass.: MIT Press, 1993.

Korsgaard, Christine. ‘The Sources of Normativity’. In *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard, and Peter Railton. New York: Oxford University Press, 1997.

Korsgaard, Christine M. ‘Skepticism about Practical Reason’. *The Journal of Philosophy* 83, no. 1 (1986): 5–25.

Kotowicz, Zbigniew. ‘The Strange Case of Phineas Gage’. *History of the Human Sciences* 20, no. 1 (2007): 115–31. https://doi.org/10.1177/0952695106075178.

Kuhn, Steven. ‘Prisoner’s Dilemma’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, n.d. https://plato.stanford.edu/archives/fall2014/entries/prisoner-dilemma/.

Kummer, Hans. ‘Ways Beyond Appearances’. *Journal of Consciousness Studies* 7, no. 1–2 (2000): 48–51.

LaFollette, H., and I. Persson. *The Blackwell Guide to Ethical Theory*. Wiley, 2013.

Lewontin, Richard. ‘The Evolution of Cognition: Questions We Will Never Answer’. In *An Invitation to Cognitive Science*, edited by Saul Sternberg and Don L. Scarborough, 2nd ed., 4:107–32. Cambridge, MA: MIT Press, 1998.

Lewontin, Richard, and Stephen J. Gould. ‘The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme’. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205, no. 1161 (1979): 581–98.

Lloyd, Elizabeth. ‘Units and Levels of Selection: An Anatomy of the Units of Selection Debates’. In *Thinking about Evolution: Historical, Philosophical, and Political Perspectives*, edited by R.S. Singh, C.B. Krimbas, R.C. Lewontin, R. Shankar, D.B. Paul, and J. Beatty, 267–91. Cambridge University Press, 2001.

Ludwig, Kirk. ‘The Epistemology of Thought Experiments : First Person versus Third Person Approaches’. In *Midwest Studies in Philosophy*, edited by Peter A. French and Howard K. Wettstein, 31:128–59. Blackwell, 2007.

Mackie, John L. *Ethics: Inventing Right and Wrong*. Penguin, 1990.

Maienshein, Jane, and Michael Ruse. *Biology and the Foundation of Ethics*. Cambridge, UK; New York: Cambridge University Press, 1999.

Mathews, K. E., and L. K. Cannon. 'Environmental Noise Level as a Determinant of Helping Behaviour'. *Journal of Personality and Social Psychology* 32, no. 4 (1975): 571–77.

McDowell, John. 'Might There Be External Reasons?' In *World, Mind, and Ethics*, edited by J. E. J. Altham and Ross Harrison. Cambridge University Press, 1995.

———. 'Values and Secondary Qualities'. In *Morality and Objectivity*, edited by Ted Honderich, 110–29. Routledge, 1985.

Melden, A. I. 'Why Be Moral?' *The Journal of Philosophy* 45, no. 17 (1948): 449–56.

Mikhail, J. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press, 2011.

Milgram, Stanley. 'Behavioural Study of Obedience'. *Journal of Abnormal and Social Psychology* 67, no. 4 (1963): 371–78.

———. *Obedience to Authority*. New York: Harper & Row, 1974.

Moore, G. E. *Principia Ethica*. Dover Publications, 2012.

Moretto, Giovanna, Elisabetta Ladavas, Flavia Mattioli, and Giuseppe Di Pellegrino. *A Psychophysiological Investigation of Moral Judgment after Ventromedial Prefrontal Damage*. Vol. 22, 2009. https://doi.org/10.1162/jocn.2009.21367.

Nado, Jennifer Ellen, Daniel Kelly, and Stephen Stich. 'Moral Judgment'. In *The Routledge Companion to Philosophy of Psychology*, edited by John Symons and Paco Calvo. Routledge, 2009.

Nagel, Thomas. *The Possibility of Alturism*. Oxford: Clarendon Press, 1970.

———. *The View from Nowhere*. oxford university press, 1989.

Neilsen, Kai. *Why Be Moral?* New York: Prometheus Books, 1989.

Nichols, Shaun. 'How Psychopaths Threaten Moral Rationalism'. *The Monist* 85, no. 2 (2002): 285–303.

———. 'Is It Irrational to Be Amoral? How Psychopaths Threaten Moral Rationalism'. *The Monist* 85, no. 2 (2002): 285–304.

———. 'Moral Rationalism and Empirical Immunity'. *Moral Psychology* 3 (2008): 395–408.

———. 'Norms with Feeling: Towards a Psychological Account of Moral Judgment'. *Cognition* 84 (2002): 221–36.

———. *Sentimental Rules: On the Natural Foundation of Moral Judgment*. 2004: Oxford University Press, 2004.

Nielsen, Rasmus. 'Adaptationism—30 Years after Gould and Lewontin'. *Evolution* 63, no. 10 (2009): 2487–90. https://doi.org/doi:10.1111/j.1558-5646.2009.00799.x.

Nietzche, Friedrich. *The Genealogy of Morals*. Anchor Books, Doubleday & Company, 1956.

Nussbaum, Martha C. 'Aristotle on Human Nature and the Foundations of Ethics'. In *World, Mind, and Ethics*, edited by J. E. J. Altham and Ross Harrison. Cambridge University Press, 1995.

Okasha, Samir. 'Biological Altruism'. edited by Edward N. Zalta. Vol. Summer 2005. The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/sum2005/entries/altruism-biological/, 2005.

———. 'Recent Work on the Levels of Selection Problem'. *The Human Nature Review* 3 (2003): 349–56.

Olson, Robert G. *The Morality of Self-Interest*. New York: Harcourt, Brace & World, 1965.

Pearson, Christopher. 'How-Possibly Explanation in Biology: Lessons from Wilhelm His's "Simple Experiments" Models'. *Philosophy, Theory, and Practice in Biology* 10, no. 4 (2018).

Pfennig, David W. 'Kin Recognition'. In *Encyclopedia of Evolution*, edited by M. Pagel, 592–95. Oxford: Oxford University Press, 2002.

Pigden, Charles R. 'Logic and the Autonomy of Ethics'. *Australasian Journal of Philosophy* 67, no. 2 (1 June 1989): 127–51. https://doi.org/10.1080/00048408912343731.

———. 'Snare's Puzzle/Hume's Purpose: Non-Cognitivism and What Hume Was Really up to with No-Ought-from-Is'. In *Hume on Is and Ought*, edited by Charles R. Pigden. Palgrave Macmillan, 2010.

Plato. *The Republic*. Edited by Desmond Lee. Penguin Books Limited, 2007.

Polonioli, Andrea. 'New Issues for New Methods: Ethical and Editorial Challenges for an Experimental Philosophy'. *Science and Engineering Ethics* 23, no. 4 (2017): 1009–34. https://doi.org/10.1007/s11948-016-9838-2.

Prichard, Harold Arthur. 'Does Moral Philosophy Rest on a Mistake?' *Mind* 21, no. 81 (1912): 21–37.

Prior, Arthur N. 'The Autonomy of Ethics'. *Australasian Journal of Philosophy* 38, no. 3 (1960): 199–206.

Rachels, James. *Created From Animals: The Moral Implications of Darwinism*. Oxford, New York: Oxford University Press, 1990.

———. *The Elements of Moral Philosophy*. 3rd Edition. New York; London: McGraw-Hill College, 1999.

Railton, Peter. 'Darwinian Building Blocks'. *Journal of Consciousness Studies* 7, no. 1–2 (2000): 55–60.

Ratiu, Peter, and Ion-Florin Talos. 'The Tale of Phineas Gage, Digitally Remastered'. *New England Journal of Medicine* 351, no. 23 (2004): e21. https://doi.org/10.1056/NEJMicm031024.

Rawls, John. *A Theory of Justice*. Oxford: Clarendon Press, 1972.

———. 'Kantian Constructivism in Moral Theory'. In *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard, and Peter Railton. New York: Oxford University Press, 1997.

Resnik, David B. 'How-Possibly Explanations in Biology'. *Acta Biotheoretica* 39, no. 2 (1991): 141–49.

Richards, Janet Radcliffe. *Human Nature After Darwin: A Philosophical Introduction*. London, New York: Routledge, 2000.

Richerson, Peter J., and Robert Boyd. *Not by Genes Alone : How Culture Transformed Human Evolution*. Chicago: University of Chicago Press, 2005. http://www.loc.gov/catdir/toc/ecip0416/2004006601.html.

Ridley, Matt. *The Origins of Virtue: Human Instincts and the Evolution of Cooperation*. New York: Viking, 1997.

Rorty, Richard. 'Born to Be Good'. *The New York Times*, n.d. https://www.nytimes.com/2006/08/27/books/review/Rorty.t.html.

Roskies, Adina. 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy"'. *Philosophical Psychology* 16, no. 1 (1 March 2003): 51–66. https://doi.org/10.1080/0951508032000067743.

Roskies, Adina L. 'Internalism and the Evidence from Pathology'. In *Moral Psychology The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 3:191–206. Cambridge, MA: MIT Press, 2008.

Ross, Richard E., Lee Nisbett. *The Person and the Situation: Perspectives of Social Pyschology*. New York: McGraw-Hill, 1991.

Ross, W. D. *The Right and the Good*. Oxford University Press, 2002.

Ruse, Michael. *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*. Oxford, New York: Basil Blackwell, 1986.

Saver, J. L., and A. R. Damasio. 'Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy Due to Ventromedial Frontal Damage'. *Neuropsychologia* 29, no. 12 (1991): 1241–49.

Shafer-Landau, Russ. *Moral Realism: A Defence*. Oxford University Press, 2003.

Shafer-Landau, Russ, and Terence Cuneo. *Foundations of Ethics: An Anthology*. Oxford: Blackwell, 2007.

Shennan, Stephen. 'Not by Genes Alone: How Culture Transformed Human Evolution, by Peter J. Richerson and Robert Boyd'. *Biology & Philosophy* 23, no. 2 (1 March 2008): 293–99. https://doi.org/10.1007/s10539-005-9007-5.

Sigmund, Karl. *Games of Life: Explorations in Ecology Evolution and Behaviour*. New York: Oxford University Press, 1993.

Singer, Peter. *A Darwinian Left: Politics, Evolution and Cooperation*. New Haven: Yale University Press, 2000.

———. 'Famine, Affluence, and Morality'. *Philosophy and Public Affairs* 1, no. 3 (1972): 229–43.

———. *How Are We to Live?: Ethics in an Age of Self-Interest*. Amherst, New York: Prometheus Books, 1995.

Smetana, Judith G. 'Preschool Children's Conceptions of Transgressions: Effects of Varying Moral and Conventional Domain-Related Attributes.' *Developmental Psychology* 21, no. 1 (1985): 18.

Smetana, Judith G., and Judith L. Braeges. 'The Development of Toddler's Moral and Conventional Judgments'. *Merill-Palmer Quarterly* 36 (1990): 329–46.

Smith, John Maynard, and Eors Szathmary. *The Major Transitions in Evolution*. Oxford; New York: W. H. Freeman Spektrum, 1995.

Smith, Michael. *Meta-Ethics*. Aldershot: Dartmouth, 1995.

———. 'Realism'. In *A Companion to Ethics*, edited by Peter Singer. John Wiley & Sons, 2013.

———. *The Moral Problem*. Oxford; Cambridge, Mass.: Blackwell, 1994.

———. 'The Truth about Internalism'. In *Moral Psychology The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 3:191–206. Cambridge, MA: MIT Press, 2008.

Smith, Michael, David Lewis, and Mark Johnston. 'Dispositional Theories of Value'. *Proceedings of the Aristotelian Society, Supplementary Volumes* 63 (1989): 89–174.

Sober, Elliott. *Conceptual Issues in Evolutionary Biology: Third Edition*. Third Edition. Cambridge, Mass.: MIT Press, 2006.

Sober, Elliott, and David Sloan Wilson. *Unto Others: The Evolution and Pscyhology of Unselfish Behaviour*. Cambridge, Mass.: Harvard University Press, 1998.

Stent, Gunther S. *Morality as a Biological Phenomenon: Report of the Dahlem Workshop on Biology and Morals*. Life Sciences Research Reports. Berlin: Abakon-Verlagsgesellschaft, 1997.

Sterelny, Kim. 'Genes, Memes and Human History'. *Mind & Language* 19, no. 2 (1 April 2004): 249–57. https://doi.org/10.1111/j.1468-0017.2004.00257.x.

———. *The Evolution of Agency and Other Essays*. Cambridge, UK: Cambridge University Press, 2001.

———. *Thought in a Hostile World: The Evolution of Human Cognition*. Malden, MA: Blackwell, 2003.

Strandberg, Caj, and Fredrik Björklund. 'Is Moral Internalism Supported by Folk Intuitions?' *Philosophical Psychology* 26, no. 3 (2013): 319–35.

Street, Sharon. 'A Darwinian Dilemma for Realist Theories of Value'. *Philosophical Studies* 127, no. 1 (2006): 109–66.

———. 'Coming to Terms with Contingency: Humean Constructivism about Practical Reason'. *Constructivism in Practical Philosophy*, 2012, 40–59.

———. 'Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About'. *Philosophical Issues* 18, no. 1 (2008): 207–28.

Tancredi, Laurence R. *Hardwired Behaviour: What Neuroscience Reveals About Morality*. New York: Cambridge University Press, 2005.

Tinbergen, N. 'On Aims and Methods of Ethology'. *Zeitschrift Für Tierpsychologie* 20, no. 4 (12 January 1963): 410–33. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x.

Todd, Peter M., and Gerd Gigerenzer. 'Précis of Simple Heuristics That Make Us Smart'. *Behavioral and Brain Sciences* 23, no. 5 (2000): 727–41. https://doi.org/10.1017/S0140525X00003447.

Toth, Nicholas, and Kathy Schick. 'Overview of Paleolithic Archaeology'. In *Handbook of Paleoanthropology: Vol I:Principles, Methods and Approaches Vol II:Primate Evolution and Human Origins Vol III:Phylogeny of Hominids*, edited by Winfried Henke and Ian Tattersall, 1–21. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-27800-6_64-4.

Troyer, John. 'Human and Other Natures'. *Journal of Consciousness Studies* 7, no. 1–2 (2000): 62–65.

Turiel, E. M. 'Morality: Its Structure, Functions, and Vagaries'. In *The Emergence of Morality in Young Children*, edited by J. Kagan and S. Lamb. University of Chicago Press, 1990.

Van Horn, J. D., A. Irimia, C. M. Torgerson, M. C. Chambers, R. Kikinis, and A. W. Toga. 'Mapping Connectivity Damage in the Case of Phineas Gage'. *PLoS One* 7, no. 5 (2012): e37454. https://doi.org/10.1371/journal.pone.0037454.

Waal, Frans B. M. de, Stephen Macedo, Josiah Ober, and Christine Korsgaard. *Primates and Philosophers: How Morality Evolved*. Princeton, N.J.: Princeton University Press, 2006.

Wallace, R. Jay. 'Moral Psychology'. In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith. Oxford; New York: Oxford University Press, 2005.

Wegner, Daniel M. *The Illusion of Conscious Will*. Cambridge, Mass.: MIT Press, 2002.

Weston, Donna R., and Elliot Turiel. 'Act–Rule Relations: Children's Concepts of Social Rules.' *Developmental Psychology* 16, no. 5 (1980): 417.

Wheatley, Thalia, and Jonathan Haidt. 'Hypnotic Disgust Makes Moral Judgments More Severe'. *Psychological Science* 16, no. 10 (2005): 780–84.

Williams, Bernard. 'Internal and External Reasons'. In *Moral Discourse and Practice*, edited by Stephen Darwall, Alan Gibbard, and Peter Railton. New York: Oxford University Press, 1997.

Wilson, David Sloan. *Evolution for Everyone: How Darwin's Theory Can Change the Way We Think about Our Lives*. New York: Delacorte Press, 2007.

Wilson, Edward O. *On Human Nature*. Cambridge; Mass., London; England: Harvard University Press, 1978.

———. *Sociobiology: The New Synthesis*. Cambridge, Mass.: Belknap Press of Harvard University Press, 1975.

Wilson, Edward O., and Michael Ruse. 'Moral Philosophy as Applied Science'. In *Conceptual Issues in Evolutionary Ethics: Third Edition*, edited by Elliot Sober. Cambridge; Mass.: MIT Press, 1986.

Young, Liane, Fiery Cushman, Ralph Adolphs, Daniel Tranel, and Marc Hauser. 'Does Emotion Mediate the Relationship Between an Action's Moral Status and Its Intentional Status? Neuropsychological Evidence'. *Journal of Cognition and Culture* 6 (1 March 2006): 291–304. https://doi.org/10.1163/156853706776931312.

Zimbardo, Philip. *The Power and Pathology of Imprisonment. Hearings before Subcommittee No. 3, of the Committee on the Judiciary, House of Representatives, Ninety-Second Congress. Congressional Record. (Serial No. 15, 25 October 1971)*. U.S. Government Printing Office, 1971.

Zimbardo, Philip, Christina Maslach, and Craig Haney. 'Reflections on the Stanford Prison Experiment: Genesis, Transformations, Consequences'. In *Obedience to Authority: Current*

Perspectives on the Milgram Paradigm, edited by Thomas Blass. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Appendix: Consolidated lessons / guidance

The table below is compiled from the guidance provided in chapter 4 and chapter 7. The numbers shown on the left relate to the diagram presented in the conclusion of the thesis.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
1	Removing some area of philosophy entirely from the context of philosophical discourse, with the hope of revolutionising it with some new insight from another field of inquiry, is unlikely to be a successful project without the provision of good reasons for that rehoming of the problem.	§3.1.1 'The metaphysics of morality' Wilson argues ethics will be revolutionised by treating it as a biological problem but does not describe how this will happen or provide any reasons for thinking it would be successful. Removed from the context of the existing discourse it is hard to see what form this revolution takes.	Understand why the problem has traditionally been considered a philosophical problem and why other methodologies might not be suitable.
2	Examine the philosophical literature to ensure that the problem that empirical research purportedly solves is a genuine problem of philosophy, and is not a seemingly similar, but philosophically uninteresting or unrelated, problem.	§3.1.2 'The problem of altruism' Wilson argues that altruism is possible, but the kind of altruism he argues is possible is a biological conception of altruism rather than a philosophically interesting moral one.	Ensure the problem you are proposing to solve is the same problem as the philosophical literature deals with.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
3	Examine the current state of philosophical debate on the topic in question, both within literature and via engagement with philosophers with expertise in the particular area. There is not much to be gained (for philosophy anyway) in re-solving problems which philosophy has already moved on from.	§3.1.2 'The problem of altruism' It is well accepted that some forms of altruism exist. If Wilson had engaged with the philosophical literature on the topic, he would have recognised that the problem he was proposing to solve was already settled.	Ensure the problem is not already considered resolved within philosophy. Philosophical peer review may be helpful in assessing this.
4	Ascertain the relationship between what you are attempting to argue and well-known philosophical rules, for example, rules about deriving an 'ought from is' or about the relationship between naturalness and goodness. If one's argument appears to be an exception to a particular rule, then examine exactly how and why it is an exception and make this explicit. Arguments should not attempt to rehash old or well-accepted positions unless they have something new to contribute to the debate or have discovered a clear problem with the received view.	§3.1.4 'Naturalness and morality' Wilson appears to leap uncritically from what is natural to what we ought to do without awareness of the existing literature or that philosophers generally consider this argument fallacious.	Find out if there are well-known reasons why the proposed argument will not work or if there are known objections in the literature that are considered conclusive.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
5	Present a full argument, including review and response from philosophers: one way to ensure an unsuccessful attempt to integrate evolutionary or biological considerations into philosophy is to have no engagement with pre-existing philosophical dialogue.	§3.1.1 'The metaphysics of morality' Wilson suggests that taking a scientific approach and "removing ethics from the hands of philosophers" will clarify the metaphysics of morality. However, he does not provide any indication of how or what this would look like. The arguments advanced here would all have benefited from engagement with the existing philosophical dialogue and ensuring they were complete enough to say something philosophically significant.	Ensure your argument is made explicitly and is sound. The premises should be clear and the logical form valid.
6	Arguments need to be complete, including establishing the premises are true and how they lead to the conclusion. Where the conclusion is a general one, it is helpful to provide specific examples that demonstrate the general point, rather than simply asserting the general conclusion without context.	§3.1.1 'Naturalness and morality' Wilson's argument that the naturalness of certain behaviours is relevant to their ethical status was incomplete, and lacked both clear premises, a clear logical form, or any specific examples of how the ethical status was established.	Provide concrete examples of the implications where possible to aid in understanding the argument.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
7	Ensure that discussion includes potential responses to likely objections based on existing philosophical theories and actual objections from other philosophers and researchers.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce integrates evolutionary considerations into the existing philosophical debate on moral error theory.	Include anticipated and real responses from philosophers to your arguments.
8	Provide the context of the argument by discussing similar arguments, whether solely philosophical or also attempting to incorporate empirical considerations. Make clear how the argument differs from previous arguments.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce compares his arguments with those of Mackie and other error theorists to show where the similarities lie and how his argument differs.	Situate your philosophical claims within the existing literature and be explicit about how your argument differs to existing arguments.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
9	Where possible, make clear the role of empirical information or empirical considerations in the argument in question.	§3.3 'Sharon Street's Darwinian Dilemma' Street discusses how our evaluative attitudes are shaped and thoroughly saturated with evolutionary forces – they would have not been recognisable as the kinds of things they are without that evolutionary influence. She is explicit about how the empirical facts constrain the possibilities for realism and the impact these have for the metaethical status of moral realism.	Be explicit as to what role empirical considerations play in your arguments.
10	Consider seriously potential and actual responses made to arguments based on both empirical and philosophical grounds and revise the position accordingly if necessary.	§3.2 'Richard Joyce's evolutionary debunking argument' Joyce has engaged in ongoing dialogue over the years and has revised his conclusions from the strong forms of moral scepticism to a more limited conclusion that moves the burden of proof to the moral realist and challenges them to come up with a plausible account of naturalism.	Be willing to revise your position based on both philosophical responses or if the empirical research turns out to be flawed.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
11	Applying epistemological standards from other disciplines can shed light on stubborn problems if done carefully. However, to do this, analysis is required to determine whether the epistemic standards are appropriate to the argument and philosophical work is required to analyse the implications.	§3.1.1 'The metaphysics of morality' Wilson attempts to apply scientific methodology to establish the existence of certain philosophically interesting phenomena – in this case altruistic behaviour. However, his analysis of the phenomena is insufficient meaning the argument targets an apparently similar but philosophically less interesting sense of the concept altruism.	Have your interpretation of the concepts reviewed by those with expertise in the relevant area of philosophy.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
12	Often when empirical considerations are applied to moral philosophy, the goals of the arguments are sweeping and revolutionary. It is easy to overlook sound and potentially noteworthy philosophical conclusions that are revealed when arguing for these sweeping or revolutionary goals. These limited and more easily established and defended conclusions can often themselves be premises or assumptions in other philosophically interesting arguments or positions that are overlooked by focusing on the more revolutionary or dramatic conclusions.	§3.2 'Richard Joyce's evolutionary debunking argument' §3.3 'Sharon Street's Darwinian dilemma' These arguments initially tried to show that moral error theory or scepticism was true and that the entire practice of ethics needed a fundamental re-evaluation. However, while these conclusions may not have been conclusively established, more limited and potentially interesting considerations about certain aspects of moral naturalism, such as the mind-independence of moral truth or the contingent nature of morality, have been uncovered as part of those arguments.	Do not overlook minor but defensible conclusions if your argument does not support your initial more consequential conclusions.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
13	Identify philosophical theses carefully and be aware of overly general theories or terms. Often talking about philosophical positions using broad terms such as 'rationalism' may be too non-specific to usefully characterise a position.	§6.2.1 'Are Nichols' rationalisms positions held by philosophers' Nichols argues against an idea he identifies as 'moral rationalism' but cites a range of divergent sources talking about similarly divergent ideas.	Be careful of philosophical terms that are so broad that they cover a range of philosophical positions that cannot all be addressed by the same arguments.
14	Asses what philosophical assumptions your argument makes. The conclusion or importance of the argument may be significantly weakened if it is dependent on philosophical theories or premises being true that are themselves controversial or that your argument has not yet established.	§6.2.2.1 'What is wrong with psychopaths?' In arguing against empirical rationalism, Nichols assumes that motivation internalism is true. If this assumption turned out to be false, he would not be able to attribute an undisturbed capacity for moral judgment making to psychopaths and his argument would not succeed.	Check that your argument does not make controversial philosophical assumptions or if it does, you will need to defend the truth of those assumptions.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
15	Evaluate carefully that the methodology of the empirical research is robust and appropriate for the philosophical argument it is used in.	§6.2.2.2 'The moral/conventional distinction' The research that introduced the moral/conventional distinction was undertaken in developmental psychology. Because of this context, schoolyard transgressions were used that may not have had the moral importance required to distinguish them from the conventional in the later research undertaken on psychopaths.	Check that the concepts used in the empirical research precisely match the required concepts in the philosophical arguments.
16	Where possible use standard names and terminology for philosophical positions to make them more easily recognisable and to better contextualise them within the literature.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Nichols' conceptual rationalism is a form of moral motivation internalism but this is not initially clear from his description of the position. In contrast, Roskies' argument is immediately clearer than Nichols' due to her identifying that her argument targets moral motivation internalism.	Adopt existing philosophical terminology to make it easier to understand and contextualise your arguments.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
17	Surveys need to be targeted to the specific desired philosophical outcomes when crafting survey questions. Concepts that you wish to distinguish between need to be carefully addressed and targeted in the surveys/questions.	§6.2.3 'Conceptual rationalism and moral motivation internalism' The questions in Nichols' surveys left a lot of room for interpretation by participants and in some cases did not even mention the target concepts explicitly, resulting in it being unclear whether participants had the concepts Nichols was investigating in mind when responding.	Where empirical results are suggestive of interesting philosophical conclusions, it may be worth replicating the experiments but explicitly target the argument in doing so.
18	Be wary of drawing conclusions from limited or uniform samples – especially if you're trying to conclude something that is supposed to apply to a whole population or the concept used by everyone.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Nichols draws conclusions that are supposed to apply to all usages of a concept from a survey of a single sample of undergraduate students at one western university.	Ensure population samples are representative and you are not generalising too far based on limited empirical results.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
19	Surveys that show mixed results should not be interpreted as conclusive support for an argument without analysis to support that conclusion. The analysis should acknowledge and explain the distribution of responses as part of the argument supporting the conclusions reached.	§6.2.3 'Conceptual rationalism and moral motivation internalism' Surveys showing a 42/58% split of responses such as responses to Strandberg and Björklund's 'psychopath' vignette, should not be interpreted as decisive or conclusive evidence. Such a result seems more likely to indicate either a lack of shared conception or a question that can be interpreted in multiple ways.	Do not automatically interpret a majority response to surveys as conclusive in the case of empirical conceptual analysis.
20	Lack of consensus may not be due to survey methodology or interpretation; an understanding of a concept not being uniformly shared or applied by differing individuals or groups may be a reality philosophy has to live with. Arguments should be open to the result that there is diversity in how a concept is used or understood.	§6.2.3 'Conceptual rationalism and moral motivation internalism' The conclusion of Strandberg and Björklund's surveys is that there is considerable variability of concepts of moral judgment and the modality of the link between moral judgment and moral motivation within philosophically untrained undergraduates. This may imply that there is no definitive or single answer to the question of whether moral motivation is a necessary feature of moral judgement.	Do not assume all concepts will all have definite meanings amenable to conceptual analysis or will be used in uniform ways both within and across sample groups.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
21	Make clear how the argument being presented differs from the previous attempts or similar arguments made in the literature and why the new approach avoids previous objections or difficulties.	§6.3 'Adina Roskies and motivation internalism' Roskies indicates how her argument differs from similar previous counter examples to motivation internalism, and addresses why some of the common objections to prior attempts are not applicable. For example, it is more difficult to argue that VM patients only make inverted commas moral judgments. Unlike the prototypical amoralist from the previous literature, VM patients do not have any reason to be deceptive, and in most cases, it is not disputed that prior to their injuries, they had a normal mastery of moral concepts.	Be explicit about how the argument presented differs from previous attempts and how it avoids prior objections.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
22	Ensure that the research methodology correctly and robustly assesses the concepts it aims to test.	§6.3 'Adina Roskies and motivation internalism' The testing methodology Roskies used to assess VM patients' motivation following moral judgments was flawed. Roskies takes Skin Conductance Response (SCR) tests as an indication of the presence of moral motivation. But this interpretation is not backed up by the intended usage of SCRs in the research she cites or the wider theoretical understanding of SCRs within psychology and neuroscience.	Check that the methodology used in the empirical research is robust and replicable.
23	Ensure that the research cited is testing the same precise concepts that the argument in philosophy requires. Often empirical research that is suggestive of philosophical conclusions requires revision and explicit targeting to be applicable to the philosophical arguments it prompts.	§6.3 'Adina Roskies and motivation internalism' Roskies cites research that used hypothetical moral scenarios. These are not sufficient for assessing motivation in response to a moral judgment. Instead the research would require first person in situ moral judgments be tested for subsequent moral motivation.	Ensure empirical research targets the precise concepts needed, and if not consider undertaking studies to target the concepts explicitly.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
24	Just because particular examples of research are unsuccessful in supporting an argument does not mean the approach is flawed if the problems with the research could be remedied or generated in more reliable ways.	§6.3 'Adina Roskies and motivation internalism' Roskies' evidence is shown not to support the argument, but the argument is still potentially sound if the evidence can be found elsewhere or further research that addresses the methodological issues is undertaken.	An unsuccessful argument usually just shows that particular argument does not succeed, not that the conclusion cannot be otherwise established.
25	Ensure that your interpretation and understanding of philosophical or conceptual distinctions or assumptions is representative of the wider theoretical understanding within psychology or the relevant field of research. If the interpretation you adopt is not representative of the wider literature, this difference in understanding should be adequately justified.	§6.3 'Adina Roskies and motivation internalism' Roskies interpretation of the deficits of VM patients suited her argument but is not representative of the general understanding of VM patients within neuropsychology which understands VM patients as having serious deficits in moral judgment making capabilities within the context of their own lives.	Ensure your interpretation of the research's results is in line with the wider understanding of those results within that field or research.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
26	Be wary of very well-known examples or case-studies that have more mythology than substance to them and have been interpreted to support conclusions or popular ideas that the evidence does not actually support.	§6.3 'Adina Roskies and motivation internalism' Phineas Gage is well known but the actual documentation of his case history is limited, and it is unclear if he fits the typical VM patient profile.	Be wary of extremely well-known case studies and anecdotes - they may be more mythology than substance.
27	It's important that the structure of the argument is clear to avoid it being misinterpreted as something obviously unsound.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' The form of Doris and Stich's argument is a 'cannot implies ought not' argument, which tries to show that virtues of the kind virtue ethics recommend are impossible therefore it cannot be the case that we ought to embody those virtues. If this structure is not made explicit it is easy to misinterpret their argument as being a misguided interest in how things are instead of how they ought to be.	Ensure your argument is made explicitly and is sound. The premises should be clear and the logical form valid.

No.	Lesson / Guideline	Case study / Source	Summary of guidance used in diagram
28	Concepts that empirical research shows to be inaccurate need to be the same concept as used by the philosophical theory to make the argument stick.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' Doris and Stich's argument targets an overly simple conception of virtue ethics which is not held by any actual virtue ethicists.	Ensure your interpretation of philosophical concepts figuring in your argument are the same concepts the original research is interested in.
29	Arguments that overall are not successful can still contain important but more modest conclusions that are important and move debates along.	§6.4 'Social Psychology and Empirically based arguments against virtue ethics' Doris and Stich's argument may not result in the rejection of virtue ethics, but they do force it to adopt a more robust empirically based conception of virtue, and this is still a notable conclusion.	Do not overlook minor but defensible conclusions if your argument does not support your initial more consequential conclusions.